

# USE OF SCANNING BEAM FOR DIFFERENTIAL EVALUATION OF ADJACENT REGIONS FOR CHANGE IN REFLECTIVITY

## 5 CROSS-REFERENCE TO RELATED APPLICATION

This application is related to and incorporates by reference herein in its entirety a patent application entitled "DIFFERENTIAL EVALUATION OF ADJACENT REGIONS FOR CHANGE IN REFLECTIVITY", by the same inventors, Docket No. 008494USA that is being filed concurrently herewith.

10

## BACKGROUND

FIG. 1A shows a cross-sectional view of a MOS field effect transistor (FET) well known in the prior art. Such an MOS transistor typically includes source region 101s and drain region 101d, source extension region 102s and drain extension region 102d, channel 103, gate insulator 104, gate 105, and well 106, all of which are formed in semiconductor substrate 107. Additional doped layers may be present, but are not shown for simplicity. The source and drain regions 101s and 101d are heavily doped, typically with arsenic for n-type doping or boron for p-type doping. Doping levels are on the order of  $10^{20}$  dopant atoms per cubic centimeter. The layers for regions 101s and 101d are typically 500-700 angstroms deep. The extension regions 102s and 102d are also heavily doped, with the same type of dopant atoms as the source and drain regions 101s and 101d, but the extension regions are shallower – typically 300 to 500 angstroms deep. FIG. 1B shows the doping profiles in the vertical directions (along the arrow A in FIG. 1A).

25

Extension regions 102s and 102d provide contacts to the channel region 103. The transistor operates by applying a bias to the gate 105 while grounding the well 106. For example, suppose the regions 101s, 102s, 102d and 101d are n-type, so that the majority carriers are electrons. If a positive voltage is placed on gate 105 with respect to the channel 103, no current will flow between the gate 105 and channel 103 because of the presence of thin gate insulator 104.

30

However, the positive voltage will attract electrons to the channel region 103, creating a thin layer of electrons (called an inversion layer) that connects source extension 102s to drain extension 102d, allowing current to flow between the source and drain. When the voltage on gate 105 is removed, the inversion layer  
5 in channel 103 ceases to exist, and the source is disconnected from the drain. In this manner, the transistor can be turned on and off.

In practice, the doping profiles for the various source and drain layers 101s, 101d, 102s and 102d are not perfectly abrupt (box-like). They are usually  
10 formed by diffusion processes that may involve several thermal cycles, causing the profiles to be somewhat rounded. For example, FIG. 1B shows two profiles 112A and 112B for the source extension 112s, following the arrow A in FIG. 1A. Line 112A shows a relatively abrupt profile and line 112B shows a less abrupt profile. Such variation in abruptness (between lines 112A and 112B) may be  
15 encountered in different semiconductor wafers under fabrication because each step in the fabrication process has a certain tolerance. Variation in individual process steps or the cumulative variation of a series of process steps can cause a loss of abruptness in the profile (e.g. may go from line 112A to line 112B). In addition, junction depth and peak doping concentration may vary depending on  
20 process properties, such as, for example, variation in annealing temperature. For example, profile 112B forms a deeper diffused profile than profile 112A, but at a smaller peak doping concentration.

The performance of the transistor is affected by the final doping profile after  
25 annealing. Profile depth, peak concentration, and profile abruptness are carefully controlled because they directly contribute to short-channel effects and speed of the transistor. A deeper and less abrupt profile in the vertical direction creates a higher off-state leakage current that leads to increased power consumption. A smaller peak doping concentration increases the resistance  
30 component between the source/drain region and the transistor channel, leading to a greater voltage drop between the source 101s and drain 101d (FIG. 1A).

This voltage drop reduces the ability of the transistor to drive the next stage, reducing the speed of the circuit.

5 A wide variety of methods are available in the prior art to characterize doped layers in semiconductors. However, such methods are unable (to the Applicants' knowledge) to provide this characterization in a small measurement size of less than 10 $\mu$ m in diameter, and to do so without damage or contact to the layer being measured. These limitations require measurement on product wafers (as opposed to test or reference wafers) with doped areas formed in fine patterns, and effectively eliminate destructive methods such as Secondary Ion Mass Spectrometer (SIMS) and Scanning Capacitance Microscopy (SCM) from  
10 considerations.

Inference of doped layer properties is possible using electrical probing of transistors. However, this procedure requires physical contact of probes to a  
15 completed transistor structure. Moreover, this procedure is impractical in the middle of a process wherein the doped layers are being formed and the transistor is still incomplete. The time between performance of source/drain process steps and the first opportunity to electrically probe the completed structure can be days or weeks, greatly reducing the ability to implement real-time process control using electrical probing.  
20

Borden et. al. have described methods for measurement of junction depth in US Patents 6,323,951 and 6,049,220. Moreover, US Patents 5,966,019 and  
25 5,883,518 by Borden describe splitting of a laser beam into two parts of orthogonal polarization, and interference of reflection of the two orthogonal polarization components to form a signal indicative of semiconductor material properties. The just-described four patents are all incorporated by reference herein in their entirety as background.

30

Jiang et al. describe (in U.S. Patent 6,556,306 which is incorporated by reference herein in its entirety as background) a method for determining the

index of refraction of a thin film at a desired angular frequency. Jiang et al. disclose (in the abstract) generating an input desired-frequency pulse and an output detectable probe pulse. According to Jiang et al. a thin film is moved in and out of the path of the input pulse, creating an output pulse that alternates  
5 between a transmitted signal created when the film intercepts the input pulse path, and a reference signal, created when the sample is outside the input pulse path. The output pulse modulates the probe pulse, which is then detected with a photodetector and the difference between the transmitted signal and the reference signal is calculated. The above steps are repeated over a plurality of  
10 delay times between the input pulse and the probe pulse until a complete field waveform of the differential signal is characterized. The index of refraction is calculated by Jiang et al. as a function of a ratio between the differential signal for the thin film and the reference signal.

15 Heinrich et al. describe (in U.S. Patent 4,758,092 which is incorporated by reference herein in its entirety as background) a method and means for optical detection of charge density modulation in a semiconductor. Heinrich et al. describe passing a polarized coherent light beam onto an interferometer which establishes two polarized beams. The two polarized beams are focused on a  
20 silicon device under test with one beam focused on or near an active device and one beam providing a reference. After passing through the device under test the two beams are reflected off a metal layer and back through the device under test where they are recombined by the beam splitter. However, the charge carriers affect the index of refraction of silicon, and by modulating the electrical charge in  
25 the active device a small modulation occurs in the index of refraction. The modulation affects the phase delay of the one beam near the active device in relation to the reference beam, and hence when the beams are recombined at the beam splitter they interfere and convert the relative phase modulation into an amplitude modulation which can be detected with a photodiode.

30 Heinrich et al. also state that the position of a single optical beam can be spatially modulated over a silicon wafer surface to detect stationary charge

densities in one area relative to a reference area. The reference area may contain no charge density variation thereby giving an absolute reference value.

#### SUMMARY OF SEVERAL EMBODIMENTS OF THE INVENTION

5 A semiconductor wafer having at least two regions of different reflectivities on a front surface thereof is evaluated in accordance with the invention by (1) generating a beam of substantially monochromatic electromagnetic radiation of a predetermined wavelength  $\lambda$ , (2) illuminating one of the regions on the front surface with the beam; (3) rhythmically moving the beam and/or the wafer  
10 relative to one another, in an oscillation at a predetermined frequency  $f$  between the two regions, (4) continuously measuring intensity of a portion of the beam reflected by the illuminated region during uninterrupted rhythmic movement, by use of a photodetector sensitive to the wavelength  $\lambda$ , and (5) synchronously at the predetermined frequency  $f$  and in a phase-sensitive manner, detecting an  
15 amplitude of fluctuation (i.e. the time varying component) in a continuous analog electrical signal generated by the photodetector. The amplitude of fluctuation of the analog electrical signal is thereafter used as a measure of change in reflectivity between the two regions, e.g. in process control. In some embodiments the measurement is performed on a test structure that is fabricated  
20 at the same time as (and using the same processes as) active devices in the wafer.

In accordance with the invention, the predetermined wavelength  $\lambda$  of the beam incident on the front surface of the wafer is selected to be sufficiently short to  
25 ensure that an absorption length of the beam in the wafer is less than the thickness of the wafer. Use of a beam of absorption length less than wafer thickness in many embodiments of the invention ensures that any portion of the incident beam that may be reflected by the bottom surface is limited to a negligible percentage (e.g. less than 10%) of the total energy reflected from the  
30 wafer. Minimizing the contribution of the bottom surface in this manner additionally minimizes the contribution from one or more defects typically present

in the substrate, if they happen to be in a path to and from the bottom surface. Therefore, energy reflected by a wafer, when probed with a beam of predetermined wavelength  $\lambda$  is mostly reflected either from the front surface thereof, or from various regions formed by semiconductor wafer fabrication, thereby to improve sensitivity of the measurement to a level sufficient for real time process control.

In some embodiments, one of the two regions is selected to be a region of known properties (e.g. an undoped channel region of a transistor) and unknown properties of the other region (e.g. source region or drain region of the transistor) are determined. The unknown properties may be calculated (from analytical equations or numerical model) or looked up (using data obtained from reference samples), based on the amplitude of fluctuation and the known properties. Alternatively, the fluctuation amplitude may be used directly, to perform process control e.g. if the fluctuation amplitude falls outside a predetermined range then a process parameter used in fabrication of the wafer is changed.

A measure of change in reflectivity between doped and undoped regions (indicated by the amplitude of the fluctuation) can be used to determine one or more unknown properties in the doped region. Specifically, if three properties are not known, then three measures of change in reflectivity are obtained by use of lasers of three different wavelengths (one fluctuation amplitude is measured at each wavelength). These three measures of reflectivity change are then used to determine the values of each of three properties. Depending on the embodiment, the properties that are determined can be either dimensional properties (e.g. depth of a profile of doping concentration), or material properties (e.g. doping concentration), or a combination thereof (e.g. change in material property as a function of dimension, exemplified by slope of the profile of doping concentration).

Several embodiments of this invention require creation of a test structure in a semiconductor wafer and subsequent non-contact measurement of a change in

- reflectance across two regions of the test structure, by use of an oscillating beam. The test structure is fabricated in a production wafer simultaneously with, and using the same processes as regions in integrated circuits also being formed therein that are to be eventually diced from the wafer to form IC chips.
- 5 Regions in the test structure are made identical to (and formed by identical processes as) corresponding regions in the integrated circuits except that the test structure regions have an area sufficiently large to accommodate the diameter of a spot formed on the wafer by the beam.
- 10 Although in a number of embodiments of the type described above, a single beam is used, in other embodiments two beams that are coaxial are incident on a test structure and form concentric spots thereon. During relative motion of the type described above, both beams are moved relative to the test structure. While both beams are reflected by the test structure, one of the beams is filtered
- 15 out by an optical element present in the path of the reflected beams, and the other beam is incident on and detected by the photodetector in the above-described manner. In certain applications involving pre-anneal low-dose implants, the use of two beams enhances the reflectivity of, for example, a doped region due to creation of carriers by the additional beam. Therefore, the
- 20 difference in reflectivity measured by (i.e. the amplitude of fluctuation of the electrical signal generated by) the photodetector is larger, and any deviation therein is easier to detect. In other applications involving post-annealed samples, the use of two beams enhances the sensitivity of the measurement to other parameters of the doped layer such as oxide charge and defect densities, due to
- 25 creation of excess carriers by the additional beam.

#### BRIEF DESCRIPTION OF THE DRAWINGS

- FIG. 1A shows, in a cross-sectional diagram through a prior art semiconductor wafer, the structure of a MOS field effect transistor.
- 30 FIG. 1B shows, in graphs, the doping profile of the prior art semiconductor wafer, along arrow A of FIG. 1A through extension layer 102s.

FIG. 2A illustrates, in a flow chart, acts performed in several embodiments of the invention.

- 5    FIG. 2B illustrates, in a cross-sectional view, the relative dimensions of absorption length  $A_l$ , wafer thickness  $z_s$ , well depth  $z_w$  and depth  $z_j$  of an interface between doped region 211 and well region 216.

- 10    FIG. 2C illustrates, in a plan view, multiple positions of the incident beam (one position shown solid and four other positions shown dotted) during oscillating movement along a straight line L between the doped region 211 and the undoped region 213 of a test structure 210; also shown in FIG. 2C is an integrated circuit 220 formed simultaneous with test structure 210.

- 15    FIGs. 2D-2H illustrate, in a cross-sectional side view along line L of FIG. 2C, multiple positions of the incident beam of FIG. 2C.

- 20    FIG. 2I illustrates, in a graph, a plot of an analog electrical signal along the y-axis as a function of time t along the x-axis, wherein signal on the y-axis is generated by a photodetector sensitive to a wavelength of the beam of FIG. 2C in response to oscillating movement thereof, as illustrated in FIGs. 2D-2H.

- 25    FIG. 2J illustrates, in a perspective three dimensional view, the relative positions of various optical components in an embodiment that uses three lasers for three measurements, uses an acousto-optic beam deflector for causing relative motion between the beam and the wafer, and a lock-in amplifier that uses a frequency of oscillation of the beam deflector to measure amplitude of a signal generated by the photodetector.

- 30    FIGs. 3A and 3B illustrate various dimensions of the test structure relative to the beam incident thereon for use by a method of the type described herein.



FIG. 3C illustrates, in a graph, the continuous nature of the signal sensed by the photodetector, including the durations  $t_{\max}$  and  $t_{\min}$  for which the signal remains steady and the durations  $t_{\text{fall}}$  and  $t_{\text{rise}}$  during which the signal falls and rises respectively.

5

FIG. 3D illustrates, in a graph, an increase in amplitude of the analog fluctuating electrical signal of FIG. 3C which can happen in case of higher dopant concentration and/or deeper profile depth (e.g.  $z_j < 700$  micrometers), and/or longer wavelength of probe beam as compared to a structure which generates the signal of FIG. 3C.

10

FIG. 3E illustrates, in a graph, a reduction in duration  $t_{\max}$  during which time signal remains at its maximum value and an increase in duration  $t_{\min}$  during which time signal remains at its minimum value, which happens when beam positions 309a and 309b of FIG. 3B are not equidistant from a boundary 302 between the doped region 301 and undoped region 303 (specifically illustrated herein is the situation when the beam spends more time in the doped region than in the undoped region).

15

FIG. 3F illustrates, in a graph, the increase in durations  $t_{\text{rise}}$  and  $t_{\text{fall}}$  during which time the signal respectively rises from its minimum value to maximum value and vice versa, which happens in a situation wherein a boundary 302 between the doped region and the undoped region is less abrupt than in a structure which generates the signal of FIG. 3B.

20

25

FIG. 3G illustrates, in a graph, a reduction in amplitude of signal wherein the beam spends equal time in the doped and undoped regions but the beam scan (distance  $D_{\text{scan}}$  between positions 309a and 309b of FIG. 3A) is less than the beam diameter (distance  $B_{\text{diameter}}$ ).

30

FIG. 3H illustrates, in a side view, oscillation of a beam along a line across alternating fingers of doped and undoped regions.

FIG. 3I illustrates, in a graph, the fluctuating electrical signal generated by a photodetector when oscillating the beam as illustrated in FIG. 3H.

- 5 FIGs. 3J-3N illustrate application of an oscillating beam measurement of the type described above to different structures in various embodiments of the invention.

FIG. 4A illustrates, in a flow chart, acts for use of a measurement of the type described above to decide if a process for fabricating a semiconductor wafer is in  
10 control (i.e. generates wafers within predefined tolerances of material and dimension properties).

FIG. 4B illustrates, a block diagram, a wafer fabrication system including a measurement apparatus used in a process flow for wafer fabrication.  
15

FIG. 5A illustrates, in a cross-sectional view, a test structure for characterizing the profile of doped regions of a transistor, in one embodiment of the invention.

FIG. 5B illustrates, in a cross-sectional view, the test structure of FIG. 5A after  
20 removal of oxide mask 511 and annealing.

FIGs. 6A-6E illustrate, in a series of cross-sectional diagrams, a process of creating a test structure of the type illustrated in FIGs. 5A and 5B.

- 25 FIG. 7A and 7B illustrate the incidence and reflection of a beam of electromagnetic radiation in a direction perpendicular to an exposed surface of an undoped material (FIG. 7A) and a doped-material (FIG. 7B).

FIG. 8A shows a simplified graph of doping concentration ( $N$ ) as a function of  
30 depth ( $z$ ) from a front surface of a wafer in a test structure, which resembles actual doping profile in real samples in the same wafer.

FIG. 8B shows the derivative of the graph in FIG. 8A.

FIG. 9A shows, in theory, the reflected signal's intensity (along the y axis) as a function of depth of a doping profile for a silicon substrate material (along the x axis), when using a laser of wavelength of 980nm for the measurement, and for  
5 peak doping concentration and profile abruptness combinations of  $8 \times 10^{19}/\text{cm}^3$  and  $100 \text{ \AA}$  (for graph 900C),  $5 \times 10^{19}/\text{cm}^3$  and  $100 \text{ \AA}$  (for graph 900A), and  $8 \times 10^{19}/\text{cm}^3$  and  $200 \text{ \AA}$  (for graph 900B).

10 FIG. 9B illustrates use of three signals  $S_1$ ,  $S_2$ , and  $S_3$  measured from an unknown sample having a profile depth  $z_j$ , using three different wavelength laser beams. Curves 901, 902, and 903 are linear portions of a cosine response curve of the type shown in FIG. 9A, in the order of decreasing wavelength.

15 FIG. 9C illustrates use of a single signal to determine a profile depth  $z_j$ , from a linear portion of a response curve of the type shown in FIG. 9B.

FIG. 9D illustrates, in a three dimensional graph, signal ratio (Samplitude/Smax) plotted on the Z-axis, as a function of the peak doping concentration and the  
20 profile depth on the X and Y-axes, respectively, wherein measurements using two different wavelengths of the measuring beam are shown as two surfaces 921 and 922.

FIG. 9E illustrates, in a flow chart, acts performed in some embodiments of the  
25 invention.

FIG. 10A illustrates a screenshot of a commercially-available software program Mathcad 2001i Professional which is programmed in accordance with certain  
embodiments of the invention to solve three simultaneous nonlinear equations  
30 for the determination of the following three properties: peak ( $N_d$ ) of the dopant profile, depth ( $z_j$ ) of the dopant profile, and abruptness ( $a$ ) of the dopant profile using three amplitude measurements at three wavelengths.

FIGs. 10B and 10C illustrate, in a block diagram, two alternative embodiments of a three laser apparatus for performing the measurements used in the program of FIG. 10A.

5

FIG. 11A illustrates in an alternate embodiment probing of doped region 1101 and undoped region 1103 of a test structure by a carrier generation beam 1110 collinear with a measuring beam 1109.

10 FIG. 11B illustrates generation of excess carrier distributions 1113 and 1114 by beam 1110 in the doped layer 1101 and well region 1106, respectively.

FIG. 11C illustrates, in a plan view, multiple positions of two concentric incident beams (one position shown solid and two other positions shown dotted) during  
15 oscillation along a straight line L between the doped region and the undoped region of a test structure.

FIG. 12 shows the hardware configuration of some embodiments for evaluation of a test structure using either an oscillating stage (1218) or a beam deflector  
20 (1215) as the means for oscillating a spot along a straight line between two adjacent regions.

#### DETAILED DESCRIPTION OF SEVERAL EMBODIMENTS OF THE INVENTION

25

During fabrication of a semiconductor wafer in accordance with the invention, a test structure 210 (FIG. 2B) is also simultaneously fabricated (as per act 201 in FIG. 2A), using the same processes that are used to form integrated circuits (e.g. IC portion 220) in the wafer. The only difference between regions of the  
30 test structure 210 and regions of the integrated circuit portion 220 is in dimensions. Specifically, the test structure 210 is designed to have dimensions sufficiently large to accommodate the presence of a beam 219 of

electromagnetic radiation as described next. For example, a doped region 211 in the test structure may be a square with a side of the same size as the spot diameter. Making the doped region 211 of the test structure larger than the spot diameter accommodates error in alignment of the beam relative to the region.

5

Note that a doped region 211 in such a test structure may be made smaller than the beam spot in some embodiments (e.g. the doped region's diagonal may be same as the spot diameter), although the signal measured in such embodiments is not as sensitive to a change in properties between regions. Note also that the specific active device 220 that is being fabricated in the wafer is not a critical aspect of the invention, as long as a structure 210 to be tested is formed at the same time as and in the same manner as the active device 220 (or portions thereof).

10

15 In accordance with the invention, the above-described wafer is evaluated during semiconductor wafer fabrication by use of a laser (or other such coherent light source) to generate a beam 219 (FIG. 2B) of substantially (or even precisely) monochromatic electromagnetic radiation of a predetermined wavelength  $\lambda$  (as per act 202 in FIG. 2A) and illuminating a front surface F of the wafer using the continuously generated beam (as per act 203 in FIG. 2A). For the oscillating beam embodiments, beam 219 could be polarized or unpolarized depending on the embodiment.

20

Illumination of the front surface F in all embodiments described herein allows the incident electromagnetic radiation to immediately reach (and be reflected from) regions of interest, such as doped regions, undoped regions, metallic regions or other such regions that are being fabricated at the front surface F of the wafer, without having to pass through most of the wafer's thickness. This is in contrast to prior art illumination of the wafer's bottom surface as disclosed by Heinrich et al. in U.S. Patent 4,758,092. Illumination of the bottom surface as suggested by Heinrich et al. requires the incident beam to travel all the way through the thickness of a wafer to reach a metal layer (at the front surface) which reflects

25

30

the energy incident thereon, and the reflected energy travels back all the way through the wafer thickness a second time. The need to travel throughout the thickness of the wafer has two disadvantages (a) any defect in the substrate affects the measurement (especially since substrates are manufactured with a larger tolerance for defects), and (b) to ensure that the reflected energy can be meaningfully measured it is necessary that a probe beam of a large wavelength (e.g. 1.3  $\mu\text{m}$ ) be used, as mentioned by Heinrich et al. at column 3, lines 61-62).

The predetermined wavelength  $\lambda$  of beam 219 (FIG. 2B) in accordance with the invention is selected to be sufficiently short to ensure that an absorption length  $A_l$  of the beam in the wafer is less than the thickness  $z_s$  of the wafer. The term "absorption length" is used to indicate a length (also called "penetration depth") at which the original intensity  $E_o^2$  of beam 219 falls to  $(1/e)E_o^2$ . Therefore, more than a majority (about 63%) of the energy  $E_o^2$  incident on a front surface F of the wafer does not even reach the bottom surface B of the wafer, when absorption length is less than wafer thickness. The same is true in the reverse direction, i.e. more than a majority (about 63%) of all energy in fact reflected by bottom surface B of the wafer does not even emerge outside the front surface F of the wafer. Furthermore, the energy reflected by the bottom surface B is not equal to the energy incident thereon, because of reduction due to reflectivity of the bottom surface B.

In some embodiments with absorption length less than wafer thickness (e.g. if absorption length is 690  $\mu\text{m}$  and wafer thickness is 700  $\mu\text{m}$ ), then approximately 26% of the total energy incident on front surface F is transmitted to and becomes incident on bottom surface B, and therefore approximately 7.8% (of the total incident energy on front surface F) is reflected by bottom surface B, and therefore about 2.8% (of the total incident energy on front surface F) reaches front surface F and only about 2% (of the total incident energy on front surface F) emerges outside of the front surface F.

Use of beam 219 having absorption length  $Al$  less than wafer thickness  $z_s$  in many embodiments of the invention ensures that most of the energy reflected by the wafer, other than from the front surface  $F$ , is reflected by regions and interfaces that are physically located between the front and bottom surfaces, e.g. regions 211 and 216 (FIG. 2B) that are formed in substrate 230 during semiconductor wafer fabrication. Therefore, use of such a short wavelength beam ensures that any portion of the incident beam 219 that may be reflected by the bottom surface  $B$  is limited to a negligible percentage (e.g. less than 10%) of the total energy reflected from the wafer. For example, if the total reflected energy is 30% of the incident energy, the bottom surface contribution is limited to less than 3%. Minimizing the contribution of bottom surface  $B$  by an order of magnitude, based on appropriately selecting the wavelength (and hence the absorption length) minimizes the contribution from one or more defects typically present in the substrate, if they happen to be in a path to and from bottom surface  $B$ .

Reducing the absorption length (and hence the wavelength) provides only diminishing returns (i.e. marginal benefits) beyond a certain point which depends on a number of factors that may be different in each embodiment, such as the signal to noise ratio (SNR) and cost (e.g. of the laser). Therefore, in some embodiments, the absorption length  $Al$  is selected to be no smaller than depth  $z_j$  of a junction (or other profile) of a region formed during semiconductor wafer fabrication. Hence, any value for the absorption length in the range (junction depth  $<$  absorption length  $<$  wafer thickness) may be selected. Therefore, the maximum absorption length is  $Al_{max} < z_s$  and the minimum absorption length is  $Al_{min} > z_j$ .

In one embodiment, junction depth  $z_j$  is in a range of 500-800 angstroms and wafer thickness is 700  $\mu m$ . In this embodiment, beam 219 having a wavelength  $\lambda$  of 1.04  $\mu m$  can be selected for probing as described herein, because such a beam has an absorption length  $Al = 700 \mu m$ . As noted above, when using such

a beam, the contribution of the bottom surface to the energy being reflected from the wafer is a negligible percentage (about 2.4%) of the total energy incident on the front surface. To eliminate any contribution from the bottom surface and any intervening defect, beam 219 may be selected to be of an absorption length  $A_l$  which is less than or equal to one-half the wafer thickness, i.e. less than or equal to 350  $\mu\text{m}$ . In such a case, the energy being reflected by the wafer is primarily reflected from the front surface F and from regions of interest in the vicinity of front surface F.

For example, if beam 219 having a wavelength  $\lambda$  of 980 nm is selected for probing as described herein, its absorption length  $A_l = 100 \mu\text{m}$  is significantly less than 350  $\mu\text{m}$  which is the one-half wafer thickness. In this example, only about 4.8% of the incident energy reaches half way into the substrate 230, and only 0.2% reaches the bottom surface. The incident energy of this wavelength is therefore practically extinguished in just reaching the bottom surface, with no likelihood of any portion of the 0.2% being able to return back to the front surface and emerging therefrom. Hence, the only energy emerging from the front surface is energy that is reflected by the front surface itself and energy that is reflected from the interfaces and regions that occur close to the front surface F, such as the interface 231 between well 216 and substrate 230, and interface 232 between region 211 and well 216, as well as regions therebetween such as well 216 and region 211.

In numerous embodiments of the type described herein, beam 219 is selected to have a wavelength  $\lambda$  of anywhere in the range 400 nm to 1.04  $\mu\text{m}$ , wherein the lower limit is set by the absorption length being greater than junction depth (or profile depth) permitted by current technologies for Ultra Shallow (US) junctions, and the upper limit is set by the absorption length being less than the wafer thickness. Note that a beam of wavelength  $\lambda$  of 1.1  $\mu\text{m}$  or greater cannot be used as described herein because its absorption length of 1500  $\mu\text{m}$  is more than twice the wafer thickness of 700  $\mu\text{m}$  (noted above). Note that beam 219 can



have any power depending on the embodiment, e.g. between 5 mW and 100 mW. Note that beam 219 can be generated by a conventional laser diode, such as a Fiber Bragg Grating (FBG) stabilized AlGaAs laser with a wavelength of 980 nm and a power of 100 mW (part # 26-8052-100) available from JDS

5 Uniphase, San Jose, California.

Beam 219 (FIG. 2B) of the type described above forms a spot on a front surface F of a wafer being evaluated, and such a spot (shown at position 219A in FIG. 2C) is moved relative to the wafer (see act 204), in a physical oscillation at a

10 predetermined frequency  $f$  between two regions of a test structure, e.g. doped region 211 and undoped region 213 (FIG. 2C). In the just-described example, region 213 is formed of the same material as and is coextensive with a well region 216, and for this reason, region 213 is shown by a dashed line in FIGs. 2B and 2C. Beam 219 may be initially located at position 219A over a doped  
15 region 211 (FIG. 2C) and during act 204 the beam is moved continuously from position 219A to a symmetric position 219E over undoped region 213, transitioning through positions 219B, 219C and 219D therebetween along a straight line L. As shown in FIG. 2C, regions 211 and 213 are adjacent to one another and share a common boundary 212.

20

Note that only three intermediate positions are illustrated in FIG. 2C for clarity, although an innumerable number of such positions are present between positions 219A and 219E due to the continuous movement. Positions 219A and 219E are extreme end positions at which movement of the beam 219 is  
25 reversed. Therefore, after position 219E is reached, movement is reversed and the beam returns back to position 219A after transitioning through positions 219D, 219C and 219B. Side views of the five positions 219A-219E (FIG. 2C) are illustrated in the respective FIGs. 2D-2H.

30 Although certain embodiments require relative motion between a beam and the wafer, in alternative embodiments of the type described in reference to FIGs. 13-17, there is no relative motion. Instead, the intensity of a beam to be incident on

a region of the wafer is modulated, e.g. by use of a rotating half-wave plate and a Wollaston prism in a path of the beam towards the wafer, or by modulating a laser driver that drives a laser that in turn generates the beam. Therefore, in some embodiments that modulate the beam's intensity, there is no relative  
5 motion between the beam and the wafer.

Referring back to FIGs. 2D-2H, during relative motion between the two regions 211 and 213, intensity of a portion of the beam reflected by the illuminated region is continuously measured (as per act 203) by a photodetector sensitive to  
10 the wavelength  $\lambda$  of the incident beam. Note that the intensity of the radiation being reflected changes during the relative motion, because the difference in reflectivity at the two end positions is the same as the difference in reflectivity between the two regions. At a central position between the two end positions, the beam is exactly 50% in each of the two regions and the reflectivity should be  
15 an average of the two reflectivities. In practice, reflectivity at the central location may significantly deviate from the average due to presence of discontinuities that may be inherent in the boundary between the two regions. At any other intermediate position, the reflectivity is a weighted average of the area of the two regions illuminated by the incident beam.

20 Therefore, during relative movement between the wafer and the beam (between the two end positions 219A and 219E in FIG. 2B), a photodetector 215 that is sensitive to the portion of the beam reflected from the wafer generates a continuous analog electrical signal. The generated electrical signal has a  
25 magnitude that varies with time, at the predetermined frequency  $f$  of relative movement between the beam and the wafer. As per act 205 in FIG. 2A, an amplitude of fluctuation (i.e. the time varying component) of the electrical signal is synchronously detected at frequency  $f$  of the relative movement, e.g. by use of a lock-in amplifier.

30 Note that in all embodiments of the invention, a beam that is incident on the wafer is generated continuously, and the beam is used to continuously illuminate

the test structure. The continuous illumination of the test structure distinguishes this invention from the disclosure of Jiang et al. (in U.S. Patent 6,556,306) which appears to require the use of pulses. Applicants note that Jiang appears to use the term "desired frequency" to mean a specific range of frequencies, e.g. GHz –  
5 THz (column 1 lines 8-27 and column 3, line 66), instead of a monochromatic beam. Applicants further note that Jiang uses a pulse, and a pulse normally has a frequency spread that is a significant multiple (e.g. 3 terahertz) of the center frequency (e.g. 1 terahertz). For either or both reasons, Jiang's pulse is not monochromatic (in the same sense that this term is used herein to describe a  
10 laser-generated beam which is used in some embodiments of Applicants' invention).

The amplitude of fluctuation (i.e. the time varying component) of the electrical signal generated by the photodetector is also referred to herein as a reflectivity  
15 change measure. The reflectivity change measure indicates a difference in reflectance between the two regions, and is thereafter used (1) to obtain a measure of a change in one or more properties between the two regions (based on modeling or previously collected data); or (2) to perform process control (by comparison with predetermined limits and changing a process parameter if  
20 outside a range).

In some embodiments, the just-described measurements (at the innumerable positions between and including end positions 219A and 219E) are performed by use of at least one beam of electromagnetic radiation that is incident normal to a  
25 surface F of the wafer (FIG. 2D), and a reflected portion of this beam is detected. FIGs. 2D-2H depict a beam splitter 218 that is shown to be in-line with the incident beam 219, to indicate that specular reflection from surface F in a direction opposite to the direction of incidence is being measured (wherein both directions are along a line 214 normal to the surface F). Beam splitter 218  
30 passes a laser beam that is generated by laser 217 through to wafer 230. Beam splitter 218 splits off a portion (e.g. 50%) of the beam reflected by wafer 230 to a detector 215. For this reason, beam splitter 218 is also referred to elsewhere

herein as “detection system beam splitter”. Detection system beam splitter 218 is non-polarizing in many embodiments of the type described herein. Detector 215 is offset from the line of incidence and reflection 214 in the usual manner.

- 5 The above-described measurements are performed by the detector 215 continuously and sequentially (i.e. one after another), during relative motion between the beam 219 and the wafer 230, to generate an analog electrical signal of the type illustrated in FIG. 2I. An amplitude of fluctuation of the signal  
10 Samplitude is obtained as an average over a number of cycles N, by use of a lock-in device. The amplitude of fluctuation Samplitude (FIG. 2I) indicates a change in reflectivity between the two regions.

- Detecting the amplitude of fluctuation (instead of making an individual reflectivity measurement in a region) eliminates the need to determine a constant present in  
15 the individual reflectivity measurement (as a baseline Smin in the analog electrical signal). Specifically, an electrical signal varies between Smin and Smax as illustrated in FIG. 2I. For example, signal that is being measured by detector 215 is at its minimum Smin between time t1 and t2 (FIG. 2I) when the beam 219 is incident wholly within the doped region 211 (FIG. 2D) and signal  
20 reaches its maximum Smax between time t3 and t4 (FIG. 2I) when the beam 219 is incident wholly within the undoped region 213 (FIG. 2H). An amplitude Samplitude of fluctuation of the periodic signal is detected using a phase-sensitive lock-in technique to isolate the ac (alternating current) component as compared to a dc (direct current) component. Specifically, the difference Smax-  
25 Smin = Samplitude is measured by the lock-in device, and eliminates the need to calibrate the effect of Smin which is present as a baseline in the electrical signal S.

- Normally, the Samplitude in each of a number of cycles is the same, i.e.  
30 Samplitude1 = Samplitude2 = Samplitude1 = SamplitudeN. However, in some embodiments, Samplitude in each cycle is not individually detected and instead Samplitude is averaged (e.g. by the lock-in device) over a number of cycles N,

i.e.  $\text{Samplitude} = (\text{Samplitude1} + \text{Samplitude2} + \dots \text{Samplitudel} + \dots \text{SamplitudeN})/N$ . Depending on the signal to noise ratio (SNR), the number N can be one hundred or more (e.g. oscillate 100-200 times between the two regions) although in some embodiments that have a very high SNR the cycles  
5 may be as few in number as a handful (say 5-10 times).

As the incident laser beam 219 (FIG. 2F) crosses the boundary 212, beam 219 may encounter a slight discontinuity at boundary 212 in the form of a thin oxide or a dislocation in the semiconductor crystal. Due to the high sensitivity of this  
10 technique, this slight discontinuity at boundary 212 can create a spike in the analog electrical signal generated by photodetector 215. Such a spike may be present in signal (FIG. 2I), for example, between times t2 and t3 and also between times t4 and t5 (spikes are not shown in FIG. 2I).

15 Presence of spikes can be detected by tuning a lock-in amplifier to twice the frequency  $f$  of relative movement, and such spikes are used in some embodiments to indicate to the user a property of boundary 212. However, in most embodiments of the type described herein, the spikes are filtered out of signal which is otherwise fluctuating at the frequency  $f$  of relative motion.  
20 Specifically, in some embodiments, the spikes are treated as anomalies in signal from the photodetector, and are eliminated by applying a filter, in the form of a series of electrical gates. The gates are synchronized to twice the frequency  $f$  of relative movement, and the filtered signal is thereafter provided to the lock-in amplifier (for synchronous detection at the frequency of relative motion).

25 Relative motion between beam 219 and wafer 230 of the type illustrated in FIGs. 2C-2H can be accomplished in any manner, e.g. in one embodiment an optical bench (carrying the laser and the photodetector) is moved relative to a stage (carrying the wafer) which is kept stationary, while in another embodiment the  
30 stage is moved while the bench is kept stationary, while in still another embodiment the stage and the bench are both moved. The stage and/or bench can be reciprocated using a piezoelectric actuator of the type well known in the

art. See stage 1218 attached to a piezoelectric actuator 1219 of FIG. 12 (described below) which is coupled to receive an electrical signal fluctuating at the frequency  $f$  at which the lock-in detector is operated to determine Amplitude.

5

In certain embodiments, neither a stage nor a bench is moved (relative to ground), and instead a beam that passes therebetween is scanned relative to the wafer, e.g. between positions 275A and 275B in FIG. 2J. In such  
10     embodiments, the beam can be scanned using a beam deflector 274 in a path of the beam. A beam deflector is a device that takes an incoming beam traveling in a certain direction, and causes that beam to change its direction in a periodic manner about its original un-deviated direction. Beam deflector 274 can be, for example a scanning galvanometer mirror or an acousto-optic deflector, although any other optical device that can oscillate a beam in space can be used in other  
15     embodiments. FIG. 2J illustrates a beam deflector 274 that receives a beam 275 along the y axis from a dichroic mirror 272 (after passing through a beam splitter), and mirror 272 in turn receives the beam from one of lasers 271A-271C along the x axis. Beam deflector 274 oscillates beam 275 between positions 275A and 275B that are both within the x-y plane. One example of beam  
20     deflector 274 is an acousto-optic beam deflector.

In this embodiment, optical bench 279 has a planar surface 279S that is parallel to the just-described x-y plane, and the above-described optical elements illustrated in FIG. 2J are mounted on surface 279S. Note that in this  
25     embodiment, lasers 271A-271C are mounted on surface 279S, adjacent to one another, along three positions on a line L2 that is parallel to the y axis. Moreover, mirror 272 is translatable between three locations that are opposite to the three laser positions. The three locations of mirror 272 are along another line L1 that is parallel to the yaxis and therefore parallel to line L2. Moreover,  
30     beam deflector 274 is located in a path of beam 275, along line L1. Finally, a beam splitter 276 is also located along line L1, between mirror 272 and beam deflector 274.

As noted above, there are multiple ways to move a beam across the doped and undoped regions. One method uses the wafer stage to move the wafer back-and-forth under the measuring beam while the beam and detector remain stationary. The stage is oscillated over a distance sufficiently large to cover both the doped and undoped regions; for example a scan distance of 10 $\mu$ m is sufficient in some embodiments. The frequency  $f$  of oscillation of the stage can be, for example, 100 hertz (one hundred oscillations per second).

10 The specific frequency  $f$  that is used in any given embodiment depends on a number of factors, such as the type of mechanism used to implement oscillation, the delay (also called "latency") in electronics (e.g. lock-in amplifier) used to detect an electrical signal fluctuating at frequency  $f$ , the duration available for each measurement and the number of cycles to be used therein. Note that any

15 frequency in the range of 1 Hz to 20,000 Hz can be used. For example, if the modulation frequency is 1000 Hz, and if at least 10 cycles are required by the lock-in amplifier to generate a reflectance measurement, then 10 milliseconds are required to perform each reflectance measurement. In one example, the throughput is 30 wafers per hour, or 120 seconds per wafer, with each wafer

20 having a measurement taken in at least ten regions.

Note that instead of oscillating the stage, an optical bench carrying all the optical elements can itself be oscillated, although the mass of the optical bench is at least an order of magnitude larger than the mass of a stage. Another method

25 uses an acousto-optic deflector to oscillate the laser beam slightly about the propagation axis. In most embodiments, the photodetector stays stationary since it is sufficiently large to capture the reflected radiation from the end points of oscillation. Note however, that in a few embodiments, the photodetector may also be oscillated, along with the laser beam, e.g. if the photodetector is smaller

30 than the amplitude of oscillation (i.e. the distance between the two endpoints thereof).

A sound-wave traveling inside an optical crystal, e.g. GaAs, sets-up a diffraction grating for a light beam traveling in the perpendicular direction to the sound wave. The diffraction grating deflects the light beam at an angle proportional to the frequency of the sound-wave. Scanning of the light beam is achieved either  
5 by changing the sound-wave frequency at the predetermined frequency  $f$  or by modulating the amplitude of the sound-wave at the predetermined frequency  $f$ . As a slight variation, both the stage and the acousto-optic deflector are operated simultaneously at different frequencies, and the signals at the sum and/or difference frequency is detected by the lock-in amplifier.

10 Referring to FIG. 2J, in several embodiments there are additional elements located after beam deflector 274, such as the following devices: an objective lens 283, wafer 230 which is being evaluated, and a stage 281 on which the wafer is mounted. Depending on the relative orientations of the stage 281 and  
15 the bench 279, there may be additional optical elements such as a mirror 282 therebetween. Note that stage 281 can be used to move wafer 230 relative to the oscillating beam in the x, y and z directions. Specifically, stage 281 can be used to move wafer 230 in the vertical direction along the z axis to adjust focus, and in a horizontal plane to adjust the position of a test structure relative to the  
20 incident beam from mirror 282 that in turn receives the oscillating beam from acousto-optic deflector 274 and sends it to objective lens 283.

Moreover, beam positions 275A and 275B are shown exaggerated in FIG. 2I, and it is to be understood that they are sufficiently close so that the reflected  
25 beam passes back through the same optical elements (such as objective lens 283 and beam deflector 274) and is incident on a beam splitter 276 that deflects a portion of the reflected light to a photodetector 277. Note that a signal indicative of the frequency  $f$  of oscillation of deflector 274 is supplied to/from (depending on the embodiment) a lock-in amplifier to which photodetector 277 is  
30 coupled, as illustrated by the dashed arrow 278.



Note that in most embodiments of the invention, a beam that is incident on the wafer is same as the beam that is generated by a laser. Specifically, there is no intermediate secondary source for the incident beam, which is opposite to the teaching of Jiang et al. because Jiang et al. require an emitter 14 to be excited by a pulse P1 (from an optical source 12).

Similarly, in most embodiments of the invention, the portion of the reflected beam from wafer 230 that is detected by photodetector 277 is directly converted to electrical signal indicative of the properties of wafer 230. Specifically, there is no intermediate optical element such as an electro-optic sensor that is utilized in conjunction with a second beam to facilitate the detection of the probe beam. This detection method is in contrast to the teaching of Jiang et al. because Jiang et al. require the probing pulse to be detected via its interaction with a second pulse P2 in an electro-optic sensor 32 made of a 2 mm thick <110> ZnTe crystal.

A specific laser of FIG. 2J that is to be used in a measurement of the type described herein is selected in some embodiments by appropriately positioning dichroic mirror 272 along the y axis, at a predetermined location opposite to the selected laser. In FIG. 2J, mirror 272 is illustrated as being located opposite to laser 271C, although mirror 272 may be moved (along a rail that is not shown) to another location opposite to laser 271B or to yet another location opposite to laser 271A, thereby to select the wavelength to be used in illuminating the front surface of the wafer.

Instead of moving a single mirror 272, in an alternative embodiment three mirrors (not shown) are rigidly mounted relative to the respective lasers, and each mirror is used to reflect a beam only when its respective laser is turned on, by using dichroic mirrors with different transition or cut-off wavelengths. For example, when laser 271B is in use, lasers 271A and 271C are turned off, and a mirror 272B (not shown) located opposite to laser 271B reflects the beam from laser 271B along the y axis, while a corresponding mirror 272C (FIG. 2J) in the path of such a beam simply passes this beam therethrough. In one example of such

embodiment, lasers 271A, 271B, and 271C must be arranged in the order of decreasing wavelength, i.e. laser 271A has the longest wavelength ( $\lambda_1$ ), laser 271C has the shortest wavelength ( $\lambda_3$ ), and laser 271B has a wavelength in between ( $\lambda_2$ ). And then mirror 272B is designed to reflect wavelength  $\lambda_2$  but  
5 transmit wavelength  $\lambda_1$ , whereas mirror 272C is designed to reflect wavelength  $\lambda_3$  and transmit both wavelengths  $\lambda_1$  and  $\lambda_2$ . Mirror 272A, being the first one in the series, could be chosen to reflect all wavelength.

In several embodiments of the type described above, a beam that is oscillated  
10 forms spots at positions 309a and 309b (FIG. 3A) on the wafer at two different times which occur at the two ends of beam oscillation along a straight line L. A scanning distance  $D_{scan}$  between the spot positions 309a and 309b is selected to be sufficiently large so that the spots do not overlap one another (at the extreme ends of the oscillation). Therefore, the scanning distance  $D_{scan}$  is  
15 selected to be (in these embodiments) greater than a diameter of the two spots, namely  $B_{diameter}$ . Moreover, in these embodiments line L is selected to be perpendicular to a boundary 302 between doped region 301 and undoped region 303. Note that since the undoped region 303 is co-extensive with body 306, and for this reason region 303 is shown not hatched in FIGs 3A and 3B.

20 The width and length of doped region 301 of a test structure as shown in a plan view in FIG. 3A (see dimensions  $R_{width}$  and  $R_{length}$ ) are selected to be sufficiently large for region 301 to completely enclose the spot at position 309a (which is formed at an extreme end of oscillation, as described in the previous  
25 paragraph). Specifically,  $R_{width}$  and  $R_{length}$  are selected to be greater than the  $B_{diameter}$ . Such a test structure is used in some embodiments wherein dimensions of regions of transistors of integrated circuits are much finer than  $B_{diameter}$ . In all other respects doped region 301 is identical to other doped regions that are normally formed as portions of integrated circuits that are to be  
30 diced from the wafer for use in the normal manner.

- For example, depth of the doped region 301 (see  $z_j$  in FIG. 3B and in FIG. 6E) is same as a corresponding depth of such integrated circuit doped regions (such as junction depth of source/drain regions or channel regions of a transistor). Note that the term "junction depth" is used to indicate the distance from the front
- 5 surface to which dopants have been implanted in case of a junction (i.e. distance of interface 304 (FIG. 3B) from front surface F) whereas the term "profile depth" is used in a more generic situation to indicate a corresponding depth for any implant or other region that may or may not form a pn junction.
- 10 Any number of test structures may be formed and evaluated to obtain a measure of corresponding properties of the regions to be evaluated (e.g. one test structure may be formed for evaluating source regions whereas another test structure may be formed for evaluating well regions). Furthermore, there may be any number of test structures in relation to the number of integrated circuits
- 15 (e.g. there may be a couple of test structures for each integrated circuit to evaluate the source and well regions therein, or a pair of test structures may be used to evaluate the source and well regions of a group of integrated circuits).

As noted above, an analog electrical signal is generated by a photodetector,

20 and this signal has a number of characteristics that depend on the various dimensions of the test structure and on the frequency of oscillation  $f$ . For example, the duration  $t_{max}$  for which the signal remains at its maximum value  $S1$  is determined by the duration for which a spot at position 309b remains completely within undoped region 303 (at any time between  $t1$  and  $t2$  in FIG.

25 3C). As soon as a portion of the beam falls across the boundary 302 (as the beam is moved in a direction from left to right along line L) into the doped region 301, signal starts dropping (at time  $t2$  in FIG. 3C).

The drop in signal continues while the beam straddles the boundary 302, until

30 the beam has completely crossed the boundary 302 (at time  $t3$ ). Thereafter, signal remains steady at its minimum value  $S2$  while the beam reaches its extreme end of oscillation at spot position 309a, reverses the direction of

oscillation to approach boundary 302 (along line L in the direction from right to left). At this time, signal starts to rise, and continues to rise while the beam straddles the boundary 302, until the beam has completely crossed the boundary 302 (at time  $t_5$ ). Then signal remains steady at the maximum value  $S_1$  while a  
5 spot at position 309b formed by the beam is completely enclosed in region 303.

Samplitude= $S_1-S_2$  indicates a difference in reflectivity between the two regions 301 and 303, and is used in some embodiments to obtain a property of the wafer. For example, the just-described reflectivity change measure (i.e.  
10 Samplitude) is compared with certain preset maximum and minimum limits on such a reflectivity change measure (which may be predetermined from a reference wafer or from a model), assuming all other properties remain the same. In such a case, an increase in Samplitude (as shown in FIG. 3C) from a structure known to be good (as shown in FIG. 3B) can indicate any of (a) higher  
15 doping concentration, (b) deeper profile depth, and (c) less abrupt doping profile. In another example, a total of three values of the reflectivity change measure (in the form of Samplitude) are obtained using each of three laser beams of different wavelengths. The three values of the reflectivity change measure are used to obtain measures of peak ( $N_d$ ) of the dopant profile, depth ( $z_j$ ) of the dopant  
20 profile, and abruptness ( $\alpha$ ) of the dopant profile. A model of these three properties that is used in some embodiments is illustrated in FIGs. 8A and 8B and described below.

The signal generated by the photodetector contains additional information other  
25 than the reflectivity change measure Samplitude. Specifically, the maximum signal value  $S_{max}$ , and the minimum signal value  $S_{min}$  are respectively indicative of the reflectance of the two regions where the measurements are made. Moreover, the duration  $t_{max}$  for which the signal stays at its maximum value, and the corresponding duration  $t_{min}$  for which the signal stays at its  
30 minimum value can also be compared with each other. As shown in FIG. 3E, if  $t_{min}$  is greater than  $t_{max}$ , the beam is spending more time in the undoped region than in the doped region (i.e. unequal beam scan). Furthermore, the duration

trise (FIG. 3F) during which the signal rises to its maximum value from its minimum values can be indicative of a property of the boundary. For example, if trise of a production wafer is greater than trise of a reference wafer, then this is an indication that the boundary between the doped region and the undoped region of the production wafer (FIG. 3F) is more diffuse than the boundary of the reference wafer (FIG. 3C).

As noted above, the beam scan distance Dscan (FIG. 3A) across the doped and undoped regions is preferably greater than the diameter Bdiameter (which is around 2 microns in one example) of a spot formed on the wafer, in order to maximize the accuracy in detection of signal Samplitude. Therefore, in one example, Dscan is 4 microns. As the scan distance is reduced in any region, the beam spends less time with its spot fully enclosed in that region, thereby to reduce the duration for which the signal remains at its maximum value or minimum value. As illustrated in FIG. 3G, when the Dscan is less than Bdiameter, the value of Samplitude is reduced (i.e. smaller than S1-S2). In such a case, the fluctuating signal has a reduced amplitude S5-S4 which in turn results in reduced sensitivity (due to truncated Samplitude) to the change in reflectivity between the regions.

Although some embodiments generate such a truncated oscillating signal (which oscillates between S4 and S5 as shown in FIG. 3G although the true maximum and minimum are at S2 and S1 respectively), most embodiments are designed to allow signal to reach its full maximum value and full minimum value (by ensuring that the beam's spot is at least momentarily enclosed fully within each region by making the regions sufficiently large). The lower limit on the scan distance Dscan depends on the smallest value of Samplitude that can still be detected and correlated to the change in reflectivity between the regions, which in turn is determined by the noise of the system. Typically, the minimum detectable value of Samplitude is set at a signal-to-noise ratio of 1.

Although in some embodiments, a beam is scanned between two adjacent regions, in other embodiments, the beam may be scanned between two regions that are separated from one another, e.g. by one or more other regions. For example, FIG. 3H illustrates the scanning of a beam between extreme positions 388a and 388d that are respectively located over an undoped region 383L and a doped region 381R, and two additional regions 381L and 383R are present therebetween. In such a case, the beam is swept through positions 388b (at which the beam is shown currently located) and 388c that are respectively located over doped region 381L and undoped region 383R.

A fluctuating electrical signal generated by a photodetector in the just-described embodiment has a period that is one-third of the period of oscillation of the beam, as illustrated in FIG. 3I. Therefore, although in certain embodiments the frequency of oscillation of the beam (also called "beam-scan frequency") is same as the frequency of fluctuation of the electrical signal (also called "signal frequency"), in other embodiments the two frequencies can have a non-identical relationship (i.e. need not be equal).

For example, scanning of the beam across one or more pairs of intervening doped and undoped lines results in a signal frequency that is an integral multiple of the beam-scan frequency (in FIG. 3I, one doped region 381L and one undoped region 383R are together considered a single pair that intervene between regions 383L and 381R). The test structure is designed with the width of the lines larger than or equal to  $B_{\text{diameter}}$ . The signal sensed by the photodetector is analyzed in some embodiments with a lock-in amplifier set to measure the  $n$ th harmonic of the beam-scan frequency. In the example illustrated in FIG. 3I,  $n=3$ . Instead of a lock-in amplifier, in an alternative embodiment a fluctuating component in photodetector's output signal may be identified by digitally sampling the signal and a computer analyzing the samples to identify amplitude at the appropriate frequency (which is based on the frequency of mechanical oscillation of the beam relative to the test structure and also based on a pattern of the test structure across which the beam is swept).

Note that in other embodiments, instead of a pair of lines, a single line of a different material (than the two regions) may intervene, in which case the electrical signal has three steps (horizontal flat regions) instead of two steps.

- 5 Although regions 301 and 303 (FIG. 3B) have been described above as being doped and undoped regions, reflectivity change across regions of other materials and/or dimensions can also be determined by acts 202-206 described above. For example, FIG. 3J illustrates measurement (by acts 202-206) of reflectivity change between a lightly doped region 303a and doped region 301. FIG. 3K
- 10 illustrates measurement (by acts 202-206) of reflectivity change between a shallow doped region 303b and doped region 301. In such embodiments, a change in the Samplitude measurement between successive wafers being fabricated is used for process control. The Samplitude measurement may be directly compared against predetermined limits, and if not within the range then a
- 15 change is made in a process control parameter (also called "process parameter") used in fabrication of regions 303b and 301. Alternatively, the unknown properties in regions 303a and 303b may be looked up using Samplitude data obtained from reference samples with known properties.
- 20 FIG. 3L illustrates measurement (by acts 202-206) of reflectivity change between undoped region 303 and a region 301a that is formed by a layer 391 of dielectric or conductive material (e.g. gate oxide or metal). Note that in the embodiment illustrated in FIG. 3L, the front surface F of the wafer is not planar, because the just-described layer 391 is formed over a planar surface 304a of the
- 25 semiconductor material that forms the substrate (substrate 307 is explicitly shown in FIG. 3N). Therefore, front surface F of the wafer is planar in region 303 of the substrate, discontinuous at boundary 302a and planar again over region 301a of the dielectric/conductive layer 391.
- 30 The vertical distance between the two planar regions 301a and 303 (in a direction of thickness of the wafer) is thickness  $z_L$  of layer 391. Surface 304a forms an interface between substrate 306 and layer 391. In such embodiments,

a change in thickness  $z_L$  of layer 391 from one wafer to another wafer during semiconductor wafer fabrication (e.g. of layer 301a) results in a corresponding change in the Samplitude measurement which is used for process control in some embodiments. Note that the thickness of layer 391 is much less than the  
5 depth of focus of the incident beam and hence the magnitude of a change in thickness  $z_L$  of layer 391 has negligible effect on beam focusing.

Furthermore, although in some embodiments the incident beam and the reflected portion are both normal (i.e. perpendicular) to the front surface of the  
10 wafer (see line 214 normal to surface F in FIGs. 2D-2H), in other embodiments, there may be an angle langle (FIG. 3M) which is other than  $90^\circ$ , between an axis 314 of the incident beam at position 309b and surface F of wafer. In such a case, a photodetector 308b is oriented at the same angle to receive the reflected portion, as will be apparent to the skilled artisan in view of this disclosure. Also,  
15 although in some embodiments the regions being evaluated are of a test structure, in other embodiments the regions of an integrated circuit are directly evaluated as illustrated in FIG. 3N, in view of this disclosure.

Whether or not regions of an integrated circuit can be directly evaluated depends  
20 on a number of factors such as the relative dimensions of the regions and a spot formed by the incident beam. The spot formed on the wafer in turn depends on, for example, the wavelength of the beam, and the focal length of an objective lens that focuses the beam on the wafer. In one embodiment of the type illustrated in FIG. 3N, the contribution of an extension region 302s (which is  
25 located between a doped region 301s and an undoped region 303) to the measurement is ignored, by filtering out any components in the analog electrical signal that are not at the frequency of relative movement (between regions 301s and 303), e.g. by use of a lock-in detector tuned to the relative movement frequency, or by using an electrical gate to block the portion of the signal that  
30 comes from region 302s.



As noted above, several embodiments do require creation of a test structure in a semiconductor wafer and subsequent non-contact measurement of change in reflectance across portions of the test structure by use of an oscillating beam. Measurement of such test structure reflectance is used, for example, to estimate  
5 the depth of a doped region in a transistor, e.g. by forming a test structure of doped regions of the same dosage as in the transistor. In several such embodiments, the estimates can be of a dimensional property (such as junction depth which is one example of profile depth) and/or a material property (such as peak dopant concentration) and/or a mixed dimensional/material property (such  
10 as the gradient of dopant concentration which is also called profile abruptness).

Estimates of the type described herein can be made by comparing an attribute of a signal measured on the test structure of a wafer under fabrication with a corresponding structure on a set reference wafers having a set of known  
15 properties forming a calibration or look-up table. As noted above, use of a test structure in a production wafer as described herein provides a convenient method for determining various such properties because one or more dimensions of the test structure need not be limited to dimensions of the transistor or other devices to be used in an integrated circuit (being formed).  
20 Several embodiments use the measurement to control ion implantation and/or annealing processes for forming such regions in the next production wafer, e.g. by changing one or more process parameters such as implanted dose, annealing temperature, and thermal exposure time.

25 Embodiments of the type described herein have numerous advantages over the prior art methods. From the physics point-of-view, the method/apparatus of certain embodiments of the invention measure the active doping profile, since the measurement is sensitive to the majority carrier distribution, which closely follows the active doping profile. Some embodiments extract the following three  
30 properties of a semiconductor wafer resulting from doping: profile depth, profile abruptness, and peak-doping concentration. From the hardware set-up and

stability point-of-view, such embodiments of the invention offer easier optical beam alignment and long-term stability over the prior art.

FIGs. 4A and 4B illustrate the use of certain embodiments in process control, to  
5 control the wafer fabrication process in a feedback loop. Specifically, one or more test structures are formed in a production wafer, e.g. simultaneous with transistor fabrication for integrated circuits, as illustrated in act 401 of FIG. 4A. The test patterns are preferably identical in doping characteristics to a transistor layer that is of interest. For example, if source and drain formation processes  
10 used to form regions 101 and 102 are to be controlled, then the same processes are used to form the test structure(s).

Such test structures may be open areas several microns on a side situated in open areas such as the streets between integrated circuits, and may be formed  
15 simultaneously with the formation of source/drain 101 and 102. In one embodiment, junction depth of doped regions of a transistor is to be estimated, and the test structure formed in act 401 requires implantation of dopant atoms and annealing, and depending on the embodiment such implantation and annealing may be performed simultaneously with implantation and annealing for  
20 formation of transistors of the wafer.

Once one or more test structures are formed, a wafer containing the test structures is aligned to a measurement system (see act 402 in FIG. 4A), followed by measurement of a signal indicative of a property of the doped region, using a  
25 non-contact probe (see act 403 in FIG. 4A). One or more of acts 402 and 403 may be repeatedly performed, e.g. for multiple test structures as illustrated by act 408, and may be interleaved with or performed simultaneously with other kinds of measurements as would be apparent to the skilled artisan. Act 403 may be performed by any method described herein, e.g. by performing acts 202-206  
30 (FIG. 2A).

Thereafter, the signals from the measurements performed in act 403 are optionally processed to convert the signals into a more meaningful quantity related to one or more properties of the doped region, such as the peak, the depth and the abruptness of a profile of the implanted dopants (see act 404 in FIG. 4A). As noted elsewhere herein, such conversion is not necessary in some embodiments wherein the raw measurements are directly used in act 405 for process control (described next). The results of the analysis in act 404 are compared with predetermined control limits (see act 405 in FIG. 4A), and if the measurements fall within the limits, fabrication of the wafer is continued (see act 407 in FIG. 4A), followed by returning to act 401 (described above) to form additional test structures in the same wafer, or in another wafer. If the measurement falls outside the predetermined control limits, a process parameter that is used in fabrication of the wafer is changed (see act 406 in FIG. 4A), and depending on the deviation the current wafer may be rejected or alternatively may be processed further.

Therefore, measurement of the test structure's properties is performed in an in-situ manner during fabrication of a wafer and in one embodiment, a measurement tool 413 (FIG. 4B) is co-located with other wafer fabrication tools, such as an annealer 414, an ion implanter 411, a patterning tool 410 and an oxide mask removal tool 412. A wafer 430 (FIG. 4B) may enter a patterning tool 410, wherein patterns associated with the source and drain extensions of to-be formed transistors, and also doped regions of one or more test structures are formed on wafer 430.

Thereafter, wafer 430 is inserted into an ion implanter 411 wherein dopant atoms are implanted to form, for example, doped region 211 (FIGs. 2B and 2C) of a test structure 210 as well as one or more doped regions 221 of one or more integrated circuits 220 in the wafer. Next, the implant mask is removed by tool 412 (FIG. 4B), and the test structures in the wafer are evaluated by measurement tool 413 as described above in reference to act 403. Therefore, measurement tool 413 contains one or more items illustrated in FIG. 2J.

In some embodiments, measurement tool 413 contains all of the items in FIG. 2J, oriented relative to one another in the manner illustrated in FIG. 2J, and in addition also contains a stage that is located adjacent to and to the right side of optical bench 279 (FIG. 2J). Measurement tool 413 of such embodiments also  
5 contains a mirror to deflect the beam by 90° (regardless of its position being at 275A or at 275B or anywhere therebetween), and an objective lens that is located between the mirror and the wafer 230 (FIG. 2J).

10 After act 403 (FIG. 4A) is completed in measurement 413 (FIG. 4B), and if the wafer is found to have the implant properties within predetermined limits, the wafer is further processed. Specifically, the wafer is annealed in a rapid thermal annealer 414, and afterwards the test structures in the wafer are evaluated again by measurement tool 415 which is similar or identical to the above-described  
15 measurement tool 413. Since measurement tool 415 is used only after the wafer has been annealed, it is also referred to as anneal measurement tool whereas tool 413 is also referred to as an implant measurement tool.

In the embodiment illustrated in FIG. 4B, measurement signals generated by  
20 units 413 and 415 are supplied on a bus 420 that is connected via connection 422 to ion implanter 411, and via connection 421 to annealer 414, thereby to provide feedback signals to these tools 411 and 414. Alternatively, or additionally, the measurement signals on bus 420 may be provided via connection 423 to a computer 416 that is mounted on a factory floor adjacent to  
25 tools 411, 412 and 414 as well as units 413 and 415 (that are also mounted to the same floor). Due to being located on the factory floor, computer 415 is also referred to herein as a factory computer. Factory computer 416 may archive the measurement signals from units 413 and 415 for later correlation to electrical performance of electronic devices being fabricated on wafer 430. If the  
30 measurements are outside the control limits, factory computer 416 can provide corrections to ion implanter 411 and annealer 414 to return the process back to the control limit. As noted above, the feedback to ion implanter 411 and

annealer 414 may also be provided directly from either or both of measurements units 413 and 415.

FIGs. 5A and 5B show in cross-sectional views one example of such a test structure (which forms only a portion of the wafer and it is understood that the remainder of the wafer contains at least one or more integrated circuits and may contain additional test structures). In the example of FIG. 5A, the test structure includes an ion-implanted region 513 (one region is shown by way of example, although more may be used depending on other considerations, such as the ability of the measurement system to align to a pattern, or space limitations within the integrated circuit die). Region 513 of the test structure typically has a width  $R_{width}$  of  $>20\mu m$ , while its length  $R_{length}$  in the direction normal to the cross section (see FIG. 3A) may also be on the order of  $20\mu m$ , but may in some cases be as large as  $500\mu m$ .

In one embodiment, an ion implanter used to form region 513 (FIG. 5A) has exactly the same energy and dose as the ion implanter used to form source/drain extension regions of a transistor that is present (as a part of the normal circuitry) in the semiconductor wafer. Ideally, the region 513 is formed in the same ion implantation step used to form the source/drain structures on the transistors, ensuring that the ion implant in the transistors and test structures are identical. The energy and dosage for a test structure may be selected to be the same as the MOS transistor for two reasons: first, it best represents the real transistor doping and second it requires no additional process steps.

Note that the dimensions of the test structure(s) are sufficiently larger than the dimensions of the regions of transistor(s), so as to enable a signal sensed from the test structure to be evaluated in the manner described herein. For example, in several embodiments, dimensions of the test structures are an order of magnitude or more larger than transistor regions because the transistor regions are normally printed much finer than the diameter of a spot formed at the surface of the wafer by a beam of illumination incident thereon during evaluation of the

type described herein. On the other hand the dimensions of the test structure regions are selected to be of the same order of magnitude as the beam diameter (and depending on the embodiment, dimensions of test regions may be slightly smaller or slightly larger than the beam diameter).

5

In the embodiment illustrated in FIG. 5A, the dopant atoms may be implanted to a very shallow depth, typically  $<500\text{\AA}$ . In this embodiment, the implant parameters are the implant specie (B, As, P, Sb, etc), the energy (0.2 to 2 keV typically), and the dose ( $1 \times 10^{14}$  to  $3 \times 10^{15}$  atoms/cm<sup>2</sup> typically). The anneal parameters are typically the temperature (on order of 1000°C), time (instantaneous to 10 sec), ramp-up rate (50 to 200°C/sec) to temperature and ramp-down time (same as ramp-up rate). After annealing, the size of the implanted region increases due to diffusion, as shown in FIG. 5B. The originally formed region 513 is shown with dashed lines as region 514a in FIG. 5B, and after annealing it becomes larger and is labeled as region 514b. Also, in FIG. 5B the ion implantation mask 511 has been removed, which was done prior to annealing.

A process for making a test structure of the type described above in reference to FIGs. 5A and 5B is illustrated in FIGs. 6A-6E, for one embodiment. First, a photoresist layer 611 is applied to the surface of substrate 606 (FIG. 6A). Next, the photoresist layer 611 is patterned by exposing and developing the resist, creating a hole 612, in the photoresist layer 611 (FIG. 6B) for the test structure. The just-described acts are also used simultaneously for the creation of one or more portions of a transistor in the silicon wafer. For example, the source and drain regions, and extensions thereof may be formed simultaneously with formation of region 613, depending on the embodiment. If so, layer 611 has holes (such as hole 602) at the locations of the to-be-formed regions of the transistors, in addition to the hole 612 required for forming the test structure. Hole 602 typically has smaller dimensions than hole 612, depending on the embodiment.

Alternatively, all of the regions of the various transistors in a wafer may be formed by acts separate and different from the just-described acts for formation of a test structure, again depending on the embodiment.

- 5 Ion implantation is applied, to form region 613 beneath the hole 612 (and beneath any additional holes 602 that may be present for the formation of transistors as noted above). Photoresist layer 611 blocks the ion implantation elsewhere (FIG. 6C). Photoresist 611 is then removed, leaving implanted region 613 in substrate 606 (FIG. 6D), and any additional regions (such as region 603)
- 10 for the transistors. Finally, the wafer is annealed, causing diffusion of the implanted region 613 resulting in an expanded doped region 614 (FIG. 6E). Any additional region 603 if present in the wafer also experiences diffusion during annealing, resulting in a correspondingly expanded doped region 604.
- 15 As noted above, to measure properties of the doped region 604, a portion of the adjacent doped region 614 of the test structure in a semiconductor wafer is illuminated with a laser beam. Initially, when the beam is incident on region 614, it reflects off the surface of the wafer because of index of refraction difference between the air and the semiconductor material. During oscillation of
- 20 the beam into an adjacent undoped area, the optical reflection occurs only at the surface of the wafer as described next.

- FIG. 7A illustrates a beam 704 incident on a front surface F of a wafer in an undoped or low doped region 701, a portion 705 of the beam 704 being reflected from surface F, and another portion 710 being transmitted into and absorbed in
- 25 from surface F, and another portion 710 being transmitted into and absorbed in region 701. The intensity of reflected beam portion 705 is given by:

$$E_s^2 = r_s^2 E_o^2 = \left( \frac{1 - (n_o + n_s)}{1 + (n_o + n_s)} \right)^2 E_o^2 \quad (1),$$

- where  $(E_o)^2$  is the intensity of incident beam 704,  $n_o$  is the refractive index of the underlying semiconductor material (usually silicon), and  $n_s = \beta N_s$  is the change in refractive index of the semiconductor material due to its background doping
- 30

concentration  $N_s$ .  $\beta$  is a constant dependent on material properties and the wavelength of the measuring beam, which will be elaborated shortly.

On the other hand, in the doped region, the optical reflection consists of an  
 5 interference between two reflection components; one component 706 (FIG. 7B) comes from the front surface F of the wafer. The other component can be either (1) component 708 which comes from interface 703 (between doped layer 702 and the underlying region 701) or (2) component 709 which comes from every location throughout doped layer 702 having doping concentration gradient.

10

FIG. 7B illustrates the reflection process in detail. For the moment, reflection component 709 is ignored; it will be incorporated into the analysis later on.

When incident beam 704 encounters the surface F of the sample, part of it is reflected as component 706, and part of it is transmitted into doped layer 702 as  
 15 beam 707. Beam 707 is reflected from interface 703 and passes through front surface F to emerge as reflected component 708.

Beam reflected components 706 and 708 interfere to form a single reflected beam with an intensity given by:

$$20 \quad E_d^2 = \left| r_{sd} + t^2 r_j e^{-i2k(n_o + n_d)z_j} \right|^2 E_o^2 \quad (2),$$

where:

$$r_{sd} = \frac{1 - [n_o + n_d]}{1 + [n_o + n_d]} \quad (3)$$

$$r_j = \frac{[n_o + n_d] - n_o}{[n_o + n_d] + n_o} \quad (4)$$

25

$$t^2 = \frac{4n_o}{(1 + n_o)^2} \quad (5)$$

$$n_d = \beta N \quad (6)$$



$$\beta = \frac{-e^2}{2m^* \omega^2 n_o \epsilon_o} \quad (7)$$

$$\omega = kc \quad (8)$$

$$k = \frac{2\pi}{\lambda_o} \quad (9),$$

$N$  and  $z_j$  are the doping concentration and thickness (junction depth) of the doped layer, respectively,  $e$  is the electron charge,  $m^*$  is the carrier effective mass in the semiconductor material,  $\epsilon_o$  is the permittivity of free space,  $c$  is the speed of light in vacuum, and  $\lambda_o$  is the wavelength of the measuring beam. The amplitude of reflection from the surface of the doped region (beam component 706) is  $r_{sd}$ ; if there is no doping, then  $n_d = n_s$ , and  $r_{sd}$  becomes the amplitude of reflection from the undoped region (beam portion 705). The amplitude of reflection from the junction (beam reflected component 708) is  $r_j$ .

Both the front surface reflection component 706 and the interface reflection component 708 from the doped region, as detailed in equations (2)-(9), contain information about the properties of the doped layer, since the free carriers contributed by the dopant atoms after annealing modify the refractive index of the semiconductor material. The surface reflection component 706 contains information about the peak doping concentration, while the interface reflection component 708 contains additional information about profile depth and abruptness of the profile. Therefore, in several embodiments, by comparing the reflectivity of the doped and undoped region, the aforementioned properties of the doped region are measured. In practice however, even with the refractive index change due to doping, the reflectivity difference between the doped and undoped areas is actually very small, only one part in  $10^4$ - $10^6$ , depending on the doping concentration. Phase-sensitive synchronous detection is employed to enhance the difference in reflectivity; this topic will be discussed below in detail.

In order to understand how doping properties are extracted from the measurements, it is necessary to elaborate on the physics beyond equations (1)-(9). In the most general case, the doping concentration  $N$  depends on depth ( $z$ ), as in the case of implanted and annealed wafers. As a result, the change in the refractive index of the doped layer arising from the added doping ( $n_d$ ) also depends on depth. Equations (3) and (4) need to be modified slightly to reflect this dependency:

$$r_{sd} = \frac{1 - [n_o + n_d(z=0)]}{1 + [n_o + n_d(z=0)]} \quad (10)$$

$$r_j = \frac{n_d(z) - n_d(z+dz)}{n_d(z) + n_d(z+dz)} \quad (11).$$

The optical reflection from the doped layer is now distributed throughout the doping profile, at any depth where the refractive index changes, as represented by component 709 in FIG. 7B. All of the minute reflections interfere with one another and with the surface reflection to form a single returning beam. When equations (5)-(11) are substituted into equation (2), it can be shown that the reflected beam intensity from the doped layer has the form:

$$E_d^2 = r_{sd}^2 E_o^2 + \frac{|\beta| r_{sd}^2 E_o^2}{n_o} \int_0^\infty \cos(2kn_o z) \frac{dN}{dz} dz \quad (12).$$

Analytical evaluation of the integral in equation (12) in some embodiments of the invention is based on an approximation for the doping distribution  $N(z)$ . FIG. 8A depicts a simplified profile 801 of doping concentration as a function of depth ( $z$ ). The graph in FIG. 8A models adequately real doping profiles in production wafers, following ion-implantation and thermal-annealing process. As shown in FIG. 8A, the doping concentration is constant and maximum near the surface ( $N_d$ ), and then it gradually drops to the background doping concentration  $N_s$ . In

most embodiments, processes used to form background doping concentration  $N_s$  are well controlled using any prior art method, and therefore, the value of  $N_s$  is known. The slope of the doping concentration profile 801 is commonly referred to as the abruptness of the profile, and it is an important property of a  
 5 semiconductor wafer, in addition to the peak concentration  $N_d$  and the profile depth  $z_j$ .

The profile depth  $z_j$  is defined as the distance from the front surface F to a location in the interface 703 (FIG. 7B) where the concentration has dropped to a  
 10 level halfway between  $N_d$  and  $N_s$  (FIG. 8A); whereas the profile abruptness is defined as the slope  $(N_d - N_s)/2a$  (see the slanted portion of profile 801 in FIG. 8A). FIG. 8B shows the derivative of this simplified doping profile 801. As seen in FIG. 8B, the slope is zero everywhere except in the range  $2a$  centered around profile depth  $z_j$ . A sharp concentration step at the surface ( $z=0$ ) is not included  
 15 in FIG. 8B because that step has been taken into account in the description for  $r_{sd}$  in equation (10), resulting in the leading term in equation (12).

The signal of interest is the change in amplitude between reflected signals from the doped and the undoped regions. Subtracting equation (1) from equation  
 20 (12), with the use of the simplified doping profile 801 in FIG. 8A, it can be shown that the signal takes on the form:

$$Samplitude = A \frac{4|\beta|(1 - n_o)E_o^2}{(1 + n_o)^3} (N_d - N_s) \left[ 1 - \frac{\sin(2kn_o a)}{2kn_o a} \cos(2kn_o z_j) \right]$$

(13)

25 where  $A$  is a wavelength dependent coefficient representing transmission loss through optical components, photodetector conversion efficiency, and amplifier gain factor. Equation (13) indicates that the measurement of Samplitude is dependent on three properties of the doping profile, namely the profile depth  $z_j$ ,  
 30 the maximum or peak doping concentration  $N_d$ , and  $a$ , which is a property

associated with profile abruptness. Note that Equation (13) includes the interference effects of reflected component 706, 708 and 709 (FIG. 7B).

FIG. 9A shows the measurement of Samplitude (normalized by dividing it with  
5 Smax, wherein Smax is the maximum signal value as shown in FIG. 2I) plotted  
along the y-axis as a function of profile depth for a silicon substrate material  
plotted along the x-axis, when probed using a beam of wavelength 980nm.  
Specifically, three graphs 900A-900C are plotted in FIG. 9A for three  
combinations of the peak doping concentration  $N_d$  and the abruptness property  $a$   
10 (see legend at the top of FIG. 9A). The abruptness property  $a$  is defined in FIG.  
8A as being one half of the distance around profile depth  $z_j$  wherein the doping  
concentration  $N_d$  changes in substantially linear manner as a function of depth.  
The peak doping concentration primarily modifies the amplitude of the cosine  
curve, while the abruptness property modifies both the amplitude and the offset.  
15 However, as clearly evident from FIG. 9A, the effect of the profile abruptness is  
quite small (as compared to the effect of a change in dopant concentration).

A relationship between signal measurement (i.e. Samplitude) and profile depth,  
such as shown in FIG. 9A, is called a correlation curve, since it can be used to  
20 convert measured signals from unknown samples to their respective profile  
depths, provided that the other two properties  $N_d$  and  $a$  are known. For  
example, as illustrated by the dashed arrows in FIG. 9A, for  $N_d = 5e19/cm^3$  and  
 $a = 100\text{\AA}$  (see curve 900A) a measured signal of 0.01 (which represents a very  
small fraction of the incident beam intensity) corresponds to a profile depth of  
25 393 $\text{\AA}$  for the unknown sample.

Although another depth of around 1000 $\text{\AA}$  could also be the result, normally the  
smaller depth is the correct answer because in most embodiments the implant  
process does not generate implants of the large depth 1000 $\text{\AA}$ , because the wafer  
30 is undergoing formation of Ultra Shallow (US) junctions. Normally, in process  
control embodiments this ambiguity is not an issue because initially the wafer

fabrication process starts with fabrication of wafers having the smaller implant depth, and any deviations therefrom are measured (and corrected as appropriate). As an alternative to using equation (13), the correlation curve can be constructed using a set of samples, called calibration or reference samples,  
5 having known doping profiles. The doping profiles are characterized using an independent method such as SIMS or spreading resistance measurement.

In embodiments wherein all three doping properties  $z_j$ ,  $N_d$ , and  $\alpha$  are unknown, multiple measuring wavelengths are utilized. The coefficients  $A$ ,  $\beta$  and  $k$  in  
10 equation (13) change with wavelength (see equations (7)-(9)), and therefore the signal is also a function of wavelength. With at least three wavelengths, three signal measurements are made, and equation (13) is solved simultaneously for the three properties  $z_j$ ,  $N_d$ , and  $\alpha$ . The background doping concentration  $N_s$  must be known; it is usually readily available from the specifications of the starting  
15 wafer material. In several implementations, the coefficient  $A$  is also known for all the wavelengths that are utilized.

Three wavelengths require three  $A$  coefficients, and they are obtained from a calibration procedure using a single calibration sample. One patterned wafer is  
20 prepared with known active doping concentration ( $N_d$ ), profile depth ( $z_j$ ) and abruptness property ( $\alpha$ ). These properties are collected from a combination of SIMS analysis and sheet resistivity measurement. The calibration wafer is measured successively using each of the three selected wavelengths of the probing beam. For each wavelength, the coefficient  $A$  is calculated using  
25 equation (13).

Alternatively, if a calibration wafer is not available, the undoped area with known background doping concentration  $N_s$  are used to calibrate the  $A$  coefficients. In this case, the signal from the undoped area is a DC signal given by:  
30

$$S_{undoped} = A_{DC} \left( \frac{1 - n_o + |\beta| N_s}{1 + n_o - |\beta| N_s} \right)^2 E_o^2 \quad (14),$$

which is derived from equation (1). The coefficient  $A_{DC}$  may be different from the  $A$  coefficient in equation (13) due to the difference in amplifier gain between the AC and the DC signals. However, if the ratio in amplifier gains is known  
 5      beforehand, from knowledge of the amplifier circuits,  $A$  is calculated from  $A_{DC}$ , since the optical transmission loss and the photodetector conversion efficiency factors are the same. The calibration proceeds in the same way as with a calibration sample, the undoped area is measured successively using each of  
 10      the three selected wavelengths, and for each wavelength, the coefficient  $A_{DC}$  is calculated using equation (14).

Once all the coefficients in equation (13) are determined, the measurement of the doping properties of an unknown wafer proceeds by sequential probing of a  
 15      test structure containing the doped region at three pre-selected wavelengths. Three values of Samplitude are measured as signals  $S_1$ ,  $S_2$ , and  $S_3$  at the three wavelengths. As depicted in FIG. 9B, linear portions 901A, 901B, and 901C of the cosine response curve corresponding to the three wavelengths indicate the values  $S_1$ ,  $S_2$ , and  $S_3$  that are measured for a given (unknown) set of profile  
 20      depth ( $z_j$ ), peak doping concentration ( $N_d$ ), and profile abruptness ( $a$ ). Note that a linear portion 901A is present in the corresponding graph 900A (FIG. 9A) between signal values of approximately 0.003 to 0.012. In FIG. 9B, only the signal dependence on  $z_j$  is shown, for the sake of simplicity, but it should be understood that the set of curves 901A, 901B, and 901C belongs to a particular  
 25      combination of doping concentration  $N_d$  and abruptness  $a$ . Other combinations of  $N_d$  and  $a$  will result in a different set of curves 901A, 901B, and 901C. Substituting signals  $S_1$ ,  $S_2$ , and  $S_3$  into equation (13) produces three simultaneous nonlinear equations which are solved in the normal manner (e.g.

by numerical analysis) to obtain values for profile depth ( $z_j$ ), peak doping concentration ( $N_d$ ), and profile abruptness ( $a$ ).

- Note that instead of solving such equations, in certain alternative embodiments
- 5 data on the behavior of Samplitude as a function of profile depth are collected ahead of time prior to evaluation of a production wafer. Specifically, the above-described method of oscillating a beam source and detector is performed on doped and undoped regions on a number of reference wafers to obtain the values (Samplitude/Smax) for a corresponding number of different profile depths
- 10  $z_j$ . A plot of the graph from such measurements is illustrated in FIG. 9C. Note that the measurements that are plotted in FIG. 9C are for a given set of all other properties and measurement conditions, including a given wavelength of the probe beam.
- 15 Thereafter, during wafer fabrication, an Samplitude measurement on a production wafer (generated by oscillating the beam source and detector) is used to look up the profile depth, and if the resulting profile depth falls outside a predetermined range, then an appropriate change is made in a fabrication process. In one such embodiment, a pre-anneal measurement is made to
- 20 determine the dosage, and only wafers that are within predetermined tolerance limits are further processed. During anneal, the profile depth and profile abruptness are both changed, although the impact on profile depth is larger than the impact on profile abruptness. Therefore, assuming that profile abruptness has not changed, any change in the measured Samplitude is treated as
- 25 indicating a corresponding change in profile depth. Note that instead of monitoring profile depths  $z_j$  at a given set of peak doping concentration  $N_d$ , and profile abruptness  $a$ , it is also possible to monitor either the concentration  $N_d$  or the abruptness  $a$  at a given set of values of the other two properties.
- 30 In another alternative embodiment, the look-up table approach is utilized to monitor variations in any of the three properties  $z_j$ ,  $N_d$ , and  $a$ , without knowing

beforehand which property is changing. A set of reference samples with known variations in all three properties  $z_j$ ,  $N_d$ , and  $\alpha$ , clustered around a center (nominal) process condition, are measured using three laser wavelengths to construct the look-up table. The data from the look-up table can be plotted, and  
5 if all three properties are used, the resulting plot is a 4-dimensional surface for the signal for each wavelength.

For simplicity in illustrating the method of this embodiment, only two properties  $z_j$  and  $N_d$  are considered next, which generate a 3D surface when plotted. FIG. 9D  
10 shows portions of such 3D surfaces for two different wavelengths of the measuring beam (since two properties only requires two wavelength measurements). In FIG. 9D, the signal ratio (Samplitude/Smax) is plotted on the Z-axis, as a function of the peak doping concentration and the profile depth on the X and Y-axes, respectively. Surfaces 921 and 922 are constructed from  
15 measuring a given set of reference samples using illumination of two respective wavelengths  $\lambda_1$  and  $\lambda_2$ , where  $\lambda_1$  is larger than  $\lambda_2$ .

The one reference sample with the nominal process condition is represented as cross 923 on the XY-plane in FIG. 9D, with concentration  $N_{dc}$  and profile depth  
20  $z_{jc}$ . This nominal reference sample produces signals S1 and S2 on surfaces 921 and 922 respectively. When there is a slight variation in the fabrication process producing for example a slight change in doping concentration by an amount  $dN$ , the process condition moves from cross 923 to cross 924 on the XY-plane. As a result, the signals shift from S1 and S2 to a new pair of signals S1a and S2b on  
25 surfaces 921 and 922, as indicated by arrows 926 and 928 respectively.

Similarly, a slight change in profile depth by an amount  $dz$  moves the process condition from cross 923 to cross 925 on the XY plane, resulting in signal shifts from S1 and S2 to a new pair of signals S1b and S2b on surfaces 921 and 922, as indicated by arrows 927 and 929, respectively.  
30



In a typical wafer fabrication process, it may not always be possible to predict which property deviates from the nominal process condition. Several embodiments of the type described herein monitor signal deviations from nominal, and identify the process parameter that causes the signal shift, provided that only one property is changing at any given time. Referring to FIG. 9D, a change in doping concentration  $dN$  shifts the nominal signals S1 and S2 to S1a and S2a at the two wavelengths  $\lambda_1$  and  $\lambda_2$ , respectively. Similarly, a change in profile depth  $dz$  shifts the signals S1 and S2 to S1b and S2b. The signal difference at the two wavelengths are (S1a-S1) and (S2a-S2) for the doping concentration shift, and (S1b-S1) and (S2b-S2) for the profile depth shift. The ratio of the difference signal at the two wavelengths, (S1a-S1)/(S2a-S2) or (S1b-S1)/(S2b-S2), identifies the process parameter and/or property of the wafer that is causing the shift, e.g. as discussed below. The look-up table, constructed by the measurements on a set of reference samples, contains the information on the size of the signal difference ratio that can be expected to enable the identification process.

FIG. 9E further illustrates a step-by-step procedure that is used in some embodiments to implement process monitoring described in the previous paragraphs. Once again for simplicity, only two properties  $z_j$  and  $N_d$  are considered in the following description. Specifically, a set of reference wafers are fabricated with known variations in  $z_j$  and  $N_d$ , clustered around a center (nominal) process condition, as illustrated in act 931 of FIG. 9E. The reference wafers are then measured at multiple wavelengths, in this case 2 wavelengths (see act 932 in FIG. 9E). A set of look-up tables is then generated, as indicated by act 933 of FIG. 9E. Examples of such look-up tables are shown in Tables 1 and 2 below, for measurements at 670nm and 980nm wavelengths.

30

zj (Å)	Signal		Change in Signal		Ratio of Signal Change
	at 670 nm	at 980 nm	at 670 nm	at 980 nm	
230	138.4	173.0	-17.7	-27.2	0.65
240	147.3	186.5	-8.8	-13.8	0.64
250	156.1	200.2	0.0	0.0	-
260	164.7	214.3	8.7	14.1	0.62
270	173.2	228.6	17.1	28.4	0.60

Table 1: Look-up table generated by a set of reference samples with variations in profile depth  $z_j$ .

5

Nd (atoms/cm <sup>3</sup> )	Signal		Change in Signal		Ratio of Signal Change
	at 670 nm	at 980 nm	at 670 nm	at 980 nm	
$9.0 \times 10^{19}$	140.4	180.2	-15.6	-20.0	0.78
$9.5 \times 10^{19}$	148.3	190.2	-7.8	-10.0	0.78
$1.0 \times 10^{20}$	156.1	200.2	0.0	0.0	-
$1.05 \times 10^{20}$	163.9	210.2	7.8	10.0	0.78
$1.10 \times 10^{20}$	171.7	220.3	15.6	20.0	0.78

Table 2: Look-up table generated by a set of reference samples with variations in peak doping concentration  $N_d$ .

10

- 15 In this particular example, the center process condition is given by a profile depth  $z_j = 250\text{Å}$ , and a peak doping concentration  $N_d = 1.0 \times 10^{20}$  atoms/cm<sup>3</sup>. In the look-up tables, the "Change in Signal" columns refer to the signal difference as a result of a property shift from the nominal condition. For example, in Table 1, the reference wafer with  $z_j = 270\text{Å}$  gives a signal of 173.2 at 670nm, which
- 20 represents a signal change of 17.1 relative to the signal at the nominal depth of  $250\text{Å}$  ( $173.2 - 156.1 = 17.1$ ). The "Ratio of Signal Change" in the look-up table refers to the difference signal at 670nm divided by the difference signal at 980nm. For example, in Table 1, measurements on the reference wafer with  $z_j = 270\text{Å}$  yield Changes in Signal of 17.1 and 28.4 at 670nm and 980nm

wavelengths, respectively, resulting in a Ratio of Signal Change of 0.6 ( $17.1/28.4 = 0.6$ ).

- As Tables 1 and 2 indicate, a small shift (within +/- 10% of the nominal value) in  $z_j$  results in a Ratio of Signal Change in the range of 0.6 – 0.65 shown in Table 1, depending on the exact amount of the shift. However, a small shift in  $N_d$  results in a fixed Ratio of Signal Change of 0.78 in Table 2, although this ratio could be changing and different in Table 2, depending on a number of factors such as signal noise, calibration, and resolution of measurements. Comparison of a Ratio of Signal Change of an unknown wafer to see if it falls within a range in Table 1 or if it falls within a range in Table 2 enables the identification of the respective property (either  $z_j$  or  $N_d$ ) that causes the signal shift, as discussed in detail next.
- Specifically, during normal fabrication of wafers (also called “production wafers”), variations in their properties (and corresponding variations in process control parameters) are determined by measuring signals (of the type described herein) using probe beams at the same two wavelengths as in the reference wafer measurements (see act 934 in FIG. 9E). Next, the change in measured signals for the production wafers are then calculated (see act 935 in FIG. 9E), followed by taking a ratio of the signal changes at the two wavelengths (see act 936 in FIG. 9E). An example of the result of such measurement and calculation is shown in Table 3 for five production wafers with various parameter/property shifts. Comparing the calculated Ratios of Signal Change with the value ranges in the two look-up tables (Tables 1 and 2) identifies the property and/or process parameter that deviates from nominal (see act 937 in FIG. 9E), the result of which is shown in the last column of Table 3.

Wafer #	Signal		Change in Signal		Ratio of Signal Change	Property causing signal shift
	at 670 nm	at 980 nm	at 670 nm	at 980 nm		
1	142.8	179.7	-13.2	-20.5	0.64	zj
2	165.4	212.2	9.4	12.0	0.78	Nd
3	166.4	217.1	10.4	16.9	0.61	zi
4	156.1	200.2	0.0	0.0	-	none
5	143.6	184.2	-12.5	-16.0	0.78	Nd

Table 3: Measurement results of five production wafers with identification of the property that has changed based on the ratio of signal change.

- 5 In addition to identifying a property that deviates from its nominal value, such a look-up table approach is also capable of estimating the amount of shift in the out of spec property by performing a data interpolation using the look-up table values (see act 938 in FIG. 9E). Using such a shift, an appropriate adjustment is made in the fabrication process (see act 939 in FIG. 9E). For example, a feed-
- 10 back loop may be implemented by incrementally changing a process control parameter during fabrication of production wafers, and finding an incremental improvement in a corresponding property. Note that act 938 is not performed in some embodiments, i.e. the precise amount of change in a property of the wafer need not be determined in embodiments wherein act 939 is performed directly
- 15 after act 937. For example, if the calculated ratios fall within a predetermined range of ratio values then a table of predetermined data (which is determined experimentally) may be used to identify the change to be made in one or more process control parameter(s).
- 20 For example, in Table 3, wafer #3 has signal values of 166.4 and 217.1 at 670nm and 980nm wavelengths, respectively. Comparing with Table 1, these signal values for wafer #3 fall between the signal values for reference wafers with profile depths of 260Å and 270Å. A simple interpolation procedure to estimate the profile depth of wafer #3 proceeds as follows:

25

$$z_j = 260 + \frac{(166.4 - 164.7)}{(173.2 - 164.7)}(270 - 260) = 262.02\text{\AA}$$

5 for the 670nm wavelength, and

$$z_j = 260 + \frac{(217.1 - 214.3)}{(228.6 - 214.3)}(270 - 260) = 261.99\text{\AA}$$

for the 980nm wavelength.

10

The estimated profile depth is therefore 262Å, which represents a deviation of 12Å from the nominal value of 250Å.

15

If more than two process parameters are involved, the method as described above can be extended by using more than two wavelengths for the measurement, and considering all the possible combinations of the signal difference ratio (or Ratio of Signal Change). Specifically, if three different process parameters may change, then three measurements are made on each reference wafer (one with each of three probe beams of different wavelengths), and three different signal changes are computed, and then three different ratios are computed.

20

25

30

Thereafter, during wafer fabrication, each production wafer is the subject of three measurements at the three wavelengths, and once again three different ratios are computed, and they are used to identify the process parameters that have changed (again by identifying whichever ratio falls within the range of ratios in a table). The measurements on each reference wafer may be made at sufficiently close intervals in the change of each process parameter, to generate a sufficiently large "cluster" of measurements centered around nominal values (of the process parameters) so as to ensure that a single process parameter change during fabrication of production wafers is easily identified by such tables.

Instead of using the values  $S_1$ ,  $S_2$ , and  $S_3$  directly for lookup, in some embodiments such values are used in any iterative numerical methods well known to those skilled in the art. Such methods may be used to solve three simultaneous equations for the three unknown variables  $z_j$ ,  $N_d$ , and  $a$ . For example, FIG. 10A shows the screenshot of a commercial software, Mathcad 2001i Professional, which is programmed to solve the three simultaneous nonlinear equations. Block 1001 defines the equation that needs solving, i.e. equation (13), and its associated components. Some definitions for the constants are not shown in FIG. 10A, but they are defined elsewhere within the program.

Block 1002 specifies the initial guess values for the unknown variables  $N_d$ ,  $z$ , and  $a$ . Block 1003 is the “solve block”, consisting of a set of constraints (block 1004), the solving function “Find” and the solutions (block 1005). Block 1004 forms the set of three simultaneous equations to be solved, with the values  $5.019 \times 10^{-3}$ ,  $4.404 \times 10^{-3}$ , and  $3.836 \times 10^{-3}$  corresponding to the signals  $S_1$ ,  $S_2$ , and  $S_3$ , respectively. Block 1005 displays the solution to the unknown variables  $N_d$ ,  $z$ , and  $a$  as  $8.028 \times 10^{19}/\text{cm}^3$ ,  $399\text{\AA}$ , and  $102.7\text{\AA}$ , respectively, for this hypothetical unknown sample.

Full text of an example of the MathCAD program shown in FIG. 10A:

Physical constants:

Index of refraction of ambient:  $n_{\text{amb}} := 1$

Index of refraction of substrate:  $n := 3.6$

dielectric constant of free space (F/m):  $\epsilon_0 := 8.854 \cdot 10^{-12}$

dielectric constant of silicon:  $\epsilon_s := 11.7$

electron charge (coulomb):  $q := 1.602 \cdot 10^{-19}$

effective electron mass (kg):  $m_e := 2.37 \cdot 10^{-31}$

speed of light (cm/s):  $c := 3 \cdot 10^{10}$

$$N_s := 3 \cdot 10^{18}$$

$$k(\lambda) := 2 \cdot \frac{\pi}{\lambda}$$

$$\omega(\lambda) := 2 \cdot \pi \cdot \frac{c}{(\lambda \cdot 10^{-8})}$$

$$\text{beta}(\lambda) := \frac{(-q^2 \cdot 10^6)}{2 \cdot [\epsilon_0 \cdot \sqrt{\epsilon_s \cdot m_e} \cdot (\omega(\lambda))^2]}$$

$$\text{sinc}(a, \lambda) := \frac{\sin(2 \cdot k(\lambda) \cdot n \cdot a)}{2 \cdot k(\lambda) \cdot n \cdot a}$$

$$S(N_d, \lambda, z, a) := \left[ 4 \cdot \text{beta}(\lambda) \cdot \frac{(1-n)}{(1+n)^3} \right] \cdot (N_d - N_s) \cdot (1 - \text{sinc}(a, \lambda) \cdot \cos(2 \cdot k(\lambda) \cdot n \cdot z))$$

$$N_d := 10^{20}$$

$$z := 500$$

$$a := 50$$

Given

$$S(N_d, 6300, z, a) = 3.094 \cdot 10^{-3}$$

$$S(N_d, 6700, z, a) = 3.414 \cdot 10^{-3}$$

$$S(N_d, 7300, z, a) = 3.836 \cdot 10^{-3}$$

$$\text{Find}(N_d, z, a) = \begin{pmatrix} 8.029 \times 10^{19} \\ 3.992 \times 10^2 \\ 1.037 \times 10^2 \end{pmatrix}$$

The hardware configuration for several embodiments of the type described above is shown in FIG. 10B. Measurement lasers 1201a, 1201b, and 1201c are semiconductor diode lasers with wavelengths that can range anywhere from 400nm to 1100nm. Typical values of the laser wavelength that may be used are  
5 405nm, 635nm, 730nm, 830nm, and 980nm. The maximum output power of the measurement lasers is around 100mW. The output beams of the lasers 1201a, 1201b, and 1201c are collimated with collimating lenses 1202a, 1202b, and 1202c, providing collimated beam 1203. Only one of the three lasers is operated at any given time. Not shown is a mechanical shutter in front of each of the  
10 lasers to block the beams when they are not in use.

A dichroic mirror 1210 (such as a partially transmissive mirror, e.g. part number 1918-b available from Dominar of Santa Clara, Calif.) is used with one of the lasers, to select a beam 1211 that is to be incident on a surface of the wafer.  
15 Specifically, dichroic mirror 1210 is mounted on a small movable platform, and moved to an appropriate location to select the wavelength of beam 1211. Beam 1211 then passes through 50:50 non-polarizing beam splitter 1212 (e.g. part number 10BC17MB.2 from Newport Corp., Irvine, CA) which is also referred to as a "detection system beam splitter" and a 90:10 beam splitter 1213 which is  
20 also referred to as a "vision system beam splitter". An example of the vision system beam splitter 1213 is available from Precision Applied Products of Fullerton, Calif., by specifying 93.3% transmission at 0.83 microns wavelength and 90% transmission at 1.48 microns wavelength. Note that the specifications for the vision system beam splitter 1213 are selected based on the wavelength  
25 of the probe beam to ensure that a majority of the power is transmitted therethrough.

Note that although a 90:10 beam splitter is used in some embodiments as being used as a vision system beam splitter, other embodiments may use beam  
30 splitters that split the beam in different ratios, such as 95:5 or 80:20. Moreover, although a 50:50 beam splitter is used in some embodiments as a detection



system beam splitter, other embodiments may use beam splitters that split the beam in different ratios, such as 60:40.

5 The power transmitted through beam splitter 1213 thereafter reaches a beam deflector 1215 which may be, for example, an acousto-optic beam deflector or a galvanometer mirror as described elsewhere herein. An oscillating beam from deflector 1215 then passes through an objective lens 1216 (such as a 100X, 0.8 NA lens made by Olympus of Tokyo Japan), and is thereafter incident on wafer 1217. Deflector 1215 receives a modulation signal from lock-in amplifier 1231  
10 via connection 1235, which causes beam 1211 to be deflected slightly in a periodic manner about the propagation axis. The periodic deflection of beam 1211 by deflector 1215 causes beam 1211 to scan along a line on wafer 1217. Lens 1216 focuses beam 1211 onto the surface of wafer 1217, which is mounted on a stage 1218. With proper alignment, beam 1211 scans across doped and  
15 undoped regions of a test structure of interest.

The portion of beam 1211 which is specularly reflected (in the direction of incidence which is normal) from wafer 1217 is recollimated by lens 1216. Beam splitter 1213 diverts 10% of the reflected beam to lens 1226 and camera 1227,  
20 which provide a system to align the beam spot on a surface of wafer 1217 to the test structure of interest. Not shown is an autofocus system that consists of a pinhole and a detector, which also uses the portion of the reflected beam diverted by beam splitter 1213. In one embodiment, an example of lens 1226 is tube lens 81845 available from--Nikon of Tokyo, Japan, and an example of  
25 camera 1227 is a CCD camera, e.g. model 85400 available from FJW Industries of Palatine, Illinois. The signal provided by camera 1227 is fed (in one embodiment) into a vision system such as model ASP-60CR-11-S available from Cognex Corporation, Boston, Mass.

30 At least a portion of the reflected beam reaches a detector 1223, an example of which is a silicon photodiode (although a phototransistor is used in alternative embodiments). In one embodiment, detector 1223 is a photocell (such as a

silicon photodiode, e.g. PIN-44DP from UDT Sensors, Inc. of Hawthorne, CA, USA) that converts the incident electromagnetic radiation into a current. The current from detector 1223 is converted to a voltage using transimpedance amplifier 1230, the output of which goes to lock-in amplifier 1231. Note that the  
5 signal from detector 1223 to lock-in amplifier 1231 is an analog continuous electrical signal. The output of lock-in amplifier 1231 goes to a digital computer, which receives the signal and presents it (or a value derived therefrom as described herein) to the user or to other data collection systems. Lock-in  
10 amplifier 1231 includes a frequency reference that is used to modulate deflection of the beam by deflector 1215 via connection 1235. As noted above, the lock-in amplifier 1231 supplies to deflector 1215 a signal oscillating at a predetermined frequency that is generated by a reference oscillator included in the lock-in amplifier 1231.

15 In principle, embodiments of the type described herein can be extended to use more than three wavelengths to provide increased accuracy and/or to measure other properties. Additionally, for process monitoring purposes, measurement(s) of the type described above can be performed without any calibration to determine the actual value of a property. In several such applications, a  
20 reference signal value is initially obtained from a reference wafer having known good properties (e.g. formed by known process parameters). Monitoring of subsequently processed wafers in a production line utilizes a Statistical Process Control (SPC) chart, in a manner well known in the art, to track deviations from the reference signal. The signal deviations are directly related to actual process  
25 deviations, and one or more process limits are used in process control. Use of previously-identified process limit(s) results in manufacture of wafers within preset manufacturing tolerances.

In an alternative embodiment illustrated in FIG. 10C, the apparatus is identical to  
30 the above-described apparatus illustrated in FIG. 10B except for the following difference. Specifically, deflector 1215 of FIG. 10B is not used in the apparatus of FIG. 10C. Instead, in the apparatus of FIG. 10C, stage 1218 is movable and

is used to move wafer 1217 back and forth in a single direction at a fixed frequency. The signal to oscillate stage 1218 comes from lock-in amplifier 1231 via a connection 1237, as shown in FIG. 10C.

5    FIGs. 11A and 11B illustrate additional embodiment of this invention, that use two laser beams: a first laser beam 1109 is of the type described above (also called "probe beam"), and a second laser beam 1110 (also called "pump beam") has photon energy greater than the bandgap of the semiconductor material. The pump beam 1110 (FIG. 11B) is combined with the probe beam 1109 using a  
10    dichroic mirror (not shown in FIG. 11B) to form dual collinear beams. These two beams are thereafter moved together between doped region 1101 and undoped region 1103 of a test structure. The two beams are illustrated in FIG. 11A at positions 1109a and 1109b respectively at the end points of oscillation of the probe beam, and positions 1110a and 1110b respectively for the end points of  
15    oscillation of the pump beam.

As depicted in FIG. 11C, the pump beam has a slightly different (in this example slightly larger) spot size of diameter  $B_{\text{pump}}$  than the probe beam's diameter  $B_{\text{probe}}$ . FIG. 11B illustrates in more detail the effect of using the pump beam  
20    1109 in addition to the probe beam 1110. Portions of the incident probe beam 1109 and pump beam 1110 are transmitted into the doped layer 1101 (with thickness  $z_j$ ) and the underlying well region 1106 as beams 1111 and 1112, respectively. The transmitted pump beam 1112 generates excess carrier distributions 1113 and 1114 inside the doped layer 1101 and the well region  
25    1106, respectively. Due to the short carrier diffusion length inside the doped layer, the excess carrier distribution 1113 decays rapidly outside the diameter of the pump beam. Inside the well region, the carrier diffusion length is much larger, resulting in a nearly flat carrier distribution 1114. In other words, in the doped region 1101 of a test structure (see FIG. 11A), the excess carrier  
30    distribution is strongly influenced by the underlying doping profile.

On the other hand, in the undoped region 1103, the excess carrier distribution is constant with depth. The difference in reflectivity of the probe beam between two regions that are illuminated by the pump beam is now a measure of the difference in excess carrier distribution between the doped and undoped regions.

5 Therefore, an analog electrical signal that is measured by a photodetector (based on reflectivity of the probe beam) is still related to the doping parameters such as profile depth and peak doping concentration. However, the signal may also be sensitive to other parameters such as oxide charge and defect densities. In other words, by using a carrier-generation beam (i.e. the pump beam), a  
10 reflectivity measurement of the probe beam is made more sensitive to other properties of the doped layer.

As indicated earlier, phase-sensitive detection of reflectivity of the probe beam (using a lock-in device) enhances a small difference in reflectivity between the  
15 doped and undoped regions. In such embodiments, the two beams together form a combined beam which is oscillated between the doped and undoped regions in a periodic manner at a fixed frequency. A reflection of the combined beam is filtered to pass only the probe beam reflected portion to a photo-detector. An analog continuous electrical signal at the output of the photo-  
20 detector is periodic, and is detected by a lock-in amplifier at the frequency of and in phase with oscillation of the combined beam. The amplitude of this electrical signal is proportional to the difference in reflectivity between the doped and undoped areas, and contains information on one or more properties of the doped layer.

25

The difference signal of interest has the exact same form as equation (13):

$$Samplitude = A \frac{4|\beta|(1-n_o)E_o^2}{(1+n_o)^3} (N_{Cs} - N_{Cd}) \left[ 1 - \frac{Sin(2kn_o a)}{2kn_o a} Cos(2kn_o z_j) \right]$$

(15),

30

except that the doping concentrations  $N_d$  and  $N_s$  in equation (13) are replaced by  $N_{Cs}$  and  $N_{Cd}$ , which refer to the excess carrier concentrations at the undoped and doped regions, respectively.

- 5 The hardware configuration for several embodiments of the type described above in reference to FIGs. 11A and 11B is shown in FIG. 12. The apparatus of FIG. 12 is similar to the above-described apparatuses of FIGs. 10B and 10C, except for certain differences which are discussed next. In addition to three measurement lasers 1201a, 1201b, and 1201c described above, the apparatus
- 10 of FIG. 12 includes a fourth laser 1205. The fourth laser 1205 is used as a carrier generation laser, an example of which is a semiconductor diode laser operating at 830nm wavelength. In this example, its maximum output power is 200mW, to generate sufficient carrier density within the samples. Laser 1205 can also be used as a measurement laser. Not shown is a mechanical shutter in
- 15 front of laser 1205 to block the beam when not in use. The output beam of laser 1205 is collimated using collimating lens 1206, providing collimated beam 1207.

The above-described beam 1203 from one of the measurement lasers 1201a-1201c is combined with beam 1207 from the carrier generation laser using

20 dichroic mirror 1210 to create combined and collinear beam 1211. As noted above, dichroic mirror 1210 is mounted on a small movable stage/platform to allow wavelength selection for the measuring beam. Combined beam 1211 of these embodiments then passes through 50:50 beam splitter 1212, 90:10 beam splitter 1213, deflector 1215, and objective lens 1216 (100X from Olympus).

25 Also as noted above, deflector 1215 receives a modulation signal from lock-in amplifier 1231 via connection 1235, which causes the combined beam 1211 to be deflected slightly in a periodic manner about the propagation axis. Lens 1216 focuses the combined beam 1211 onto the surface of wafer 1217, which is

30 mounted on a stage 1218. Although periodic deflection of combined beam 1211 is implemented in some embodiments by deflector 1215 as described above, in alternative embodiments, stage 1218 is used to move the wafer 1217 back and

forth in one direction at a fixed frequency as described above in reference to FIG. 10C. Both types of embodiments are illustrated in FIG. 12, although it is to be understood that only one of the two mechanisms for providing relative motion need be used in any given embodiment.

5

As noted above in reference to FIGs. 10B and 10C, in the apparatus of FIG. 12 as well, the reflected beam (which includes components from the probe beam and the pump beam) is recollimated by lens 1216. Beam splitter 1213 diverts 10% of the combined reflected beam to lens 1226 and camera 1227. The  
10 combined reflected beam then enters beam splitter 1212, which deflects it to an optical filter 1221 (FIG. 12).

Filter 1221 passes the light from measurement laser 1201, but blocks light from generation laser 1205. Note that such a filter is not required in the embodiments  
15 of FIGs. 10B and 10C because of their use of a single beam. The reflected beam (after being filtered by filter 1221) at the measurement wavelength reaches detector 1223, which generates the above-described analog electrical signal. Filter 1221 of some embodiments of the apparatus of FIG. 12 is mounted on an actuator that enables it to be moved out of the path of the reflected beams, so  
20 that laser 1205 may be used by itself in such a configuration to provide a measurement wavelength (e.g. with all the other lasers 1201a-1201c turned off).

The above discussion provides a description of several embodiments. Additional embodiments and variations of such embodiments will be apparent to the skilled  
25 artisan in view of this disclosure. Several such embodiments are described next.

For example, the method of some embodiments of the invention may be applied to measure the properties of pre-annealed implant wafers. In addition, the method of may be applied to measure the properties of Silicon-On-Insulator  
30 (SOI) wafers, both pre and post-annealed. The measurement may also be used to determine the difference in dielectric coating thickness in two regions.

In addition, the above description presents use of the some embodiments for process control. However, embodiments of the type described herein may also be used for process development. For example, if a development engineer wants to compare the profile abruptness that is possible with different laser  
5 anneal treatments, an Samplitude measurement of the type described herein can provide such information. For this case, the types of test structures that can be used for evaluation are expanded than some of the test structures described above, because it is no longer necessary to manufacture the test structures within a standard flow for the manufacture of integrated circuits. For example, it  
10 is possible to use a mask of narrow poly lines, then put on spacers (silicon nitride layers on the side of the poly lines, as are commonly applied to polysilicon gates in transistors), then anneal before removing the mask to capture stress effects that the spacers may introduce. Other custom adaptations of test structures for process development are limitless in possibility, but would employ the principles  
15 discussed above.

Also, as will be apparent to the skilled artisan in view of the disclosure, any of the following properties can be determined by use of one or more Samplitude measurements of the type described herein on a wafer prior to annealing: the  
20 dose of implants in a doped region, profile depth and profile abruptness. Moreover, similar measurements after anneal are used in some embodiments to determine the depth of an annealed semiconductor junction, the dose of a doped region after annealing, doping concentration after annealing, thermal exposure of the layer, and a profile of doping concentration after annealing.

25

Also, although some embodiments use a lock-in amplifier as a synchronous detector, other implementations of a synchronous detector such as boxcar  
30 averager and transient digitizer are used in several other embodiments, as will be apparent to the skilled artisan.

Moreover, instead of a Wollaston prism as discussed above, a Rochon prism can be used in alternative embodiments. An example of the Rochon prism that

can be used in embodiments of the type described herein is available from CVI Laser as part# RCHP-10-CA-670-1064, 10 degree separation, 670-1064nm wavelength.

- 5 Therefore, numerous modifications and adaptations of the embodiments described herein are encompassed by the attached claims.